ARMY RESEARCH LABORATORY

Entity Resolution Workflow Installation Process and User Guide

by Michael H. Lee

ARL-MR-0844 July 2013

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Adelphi, MD 20783-1197

ARL-MR-0844 **July 2013**

Entity Resolution Workflow Installation Process and User Guide

Michael H. Lee
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. **REPORT DATE** *(DD-MM-YYYY)* July 2013	**2. REPORT TYPE** Final	**3. DATES COVERED (From - To)** 12 April 2013

4. TITLE AND SUBTITLE Entity Resolution Workflow Installation Process and User Guide	**5a. CONTRACT NUMBER** **5b. GRANT NUMBER** **5c. PROGRAM ELEMENT NUMBER**
6. AUTHOR(S) Michael Lee	**5d. PROJECT NUMBER** R.0006163.9 **5e. TASK NUMBER** **5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U.S. Army Research Laboratory ATTN: RDRL-CII-B 2800 Powder Mill Road Adelphi MD 20783-1197	**8. PERFORMING ORGANIZATION REPORT NUMBER** ARL-MR-0844
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)** **11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

Email: michael.h.lee.civ@mail.mil

14. ABSTRACT

This report describes the installation of the U.S. Army Research Laboratory's (ARL) Entity Resolution Workflow and its dependencies (e.g., City University of New York's **[**CUNY] Entity Extractor, Global Graph) on a Hadoop Cluster. Common installation issues are addressed and solutions provided. The report also describes the operation of the Entity Resolution Workflow Web App to perform entity resolution on an English Gigaword data.

15. SUBJECT TERMS

Entity extraction, entity resolution, global graph, RELDC, CUNY

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Michael Lee
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	UU	56	**19b. TELEPHONE NUMBER** (*Include area code*) (301) 394-5608

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

INTENTIONALLY LEFT BLANK.

1. Introduction

Entity resolution, in the context of text processing and information extraction domain, refers to the process of uniquely disambiguating a specific person or an object that appears in a text. For instance, if "John Smith" appears in a document, entity resolution seeks to identify who that "John Smith" specifically refers to from available choices in a database. This report describes the setup and configuration of the U.S. Army Research Laboratory's (ARL) software implementation of an entity resolution algorithm called Relationship-based Data Cleaning (RelDC), which "systematically exploits not only features but also relationships among entities for the purpose of disambiguation. (The main concept is that) RelDC views the database as a graph of entities that are linked to each other via relationships. It first utilizes a feature-based method to identify a set of candidate entities (choices) for a reference to be disambiguated. Graph theoretic techniques are then used to discover and analyze relationships that exist between the entity containing the reference and the set of candidates."[*] In order to demonstrate the RelDC entity resolution algorithm in an intuitive and seamless way, ARL developed an Entity Resolution Workflow (ERW). The ERW is a Hadoop application that integrates all components (entity extractor, candidate identification, entity base, etc.) required to run the RelDC algorithm. The user operates the ERW with a mouse in a Web application, so no typing is required. Once ERW is correctly configured, a user simply needs to point to an electronic document and it will automatically identify all entities in that document and disambiguate each entities. ERW hides the complexities and many intermediate steps required to set up a RelDC process. Without ERW, a user would have to manually find all entities in a document, find potential candidates of each entity, send the candidates to RelDC, and display the results to the user.

In addition to ease of use, ERW was designed to be modular so that each component can be swapped for a different implementation at a later time. For example, the first component invoked in ERW is the entity extractor. "Entity extraction (also known as named-entity recognition or entity identification) is the process of identifying tokens (typically nouns or noun phrases) in text and labeling them with predefined categories (such as person, location, organization, etc.)"[†] ERW is shipped with an entity extractor called English Information Extractor (ENIE) developed by City University of New York's (CUNY) BLENDER lab[‡]. The user is not required to use CUNY's entity extractor. There are many entity extractors available to the public that can be easily swapped into ERW if an application programming interface (API) is provided for the

[*] Kalashnikov, Dmitri; Mehrotra, Sharad. Exploiting relationships for domain-independent data cleaning, Computer Science Department, University of California, Irvine, 2004.

[†] Lee, Michael; Winkler, Robert. An Examination of Training Effects on Entity Extraction Systems in Manually Translated Iraqi Military Documents, Adelphi, Maryland, 2012.

[‡] CUNY BLENDER. http://nlp.cs.qc.cuny.edu/ (accessed June 2013).

alternate entity extractor. Another entity extractor called A Nearly-New Information Extraction System (ANNIE)[§] was successfully integrated during testing to demonstrate the interchangeability of ERW. Other components (e.g., candidate identification, entity base, result display interface) are swappable as well.

The first half of this report describes the process of installing ERW components on a Hadoop cluster. A typical Hadoop cluster consists of a primary name node, a secondary name node, and many data nodes. (For the sake of simplicity, this report describes the installation process on a Hadoop cluster with a primary name node and a single data node.) From an ERW perspective, magnitude of a Hadoop cluster is irrelevant and a Hadoop cluster with one data node is functionally equivalent to a Hadoop cluster with many data nodes. The user is free to install as many data nodes as desired, but must duplicate the instructions on all data nodes. The name node is referenced as "ERDP-NN" and the data node is referenced as "ERDP-01" for the remainder of this report.

Both ERDP-NN and ERDP-01 are running Ubuntu 10.04 LTS, and ERW is configured for a user called "arl". Due to permission issues, all ERW components will be installed under the arl user's home directory, and ERW components must be run as the arl user. The installation process should be followed as the arl user, and some parts of the installation will require executing "sudo" for commands that will make system-wide change.

A third server, referenced as "ERDP-Win7," is used for hosting Global Graph and its components. Global Graph data are accessed by the ERW components via Representational State Transfer (REST) Services running on Apache Tomcat. ERDP-Win7 is specifically running a Windows operating system (OS) to facilitate Global Graph database installation scripts.

The ERW user guide section is provided after the installation description. The user guide describes how to perform primary functions on the ERW Web Application: selecting a news article, running entity extraction, running candidate identification, running entity resolution, and viewing details on specific resolved entities.

2. Quick Start

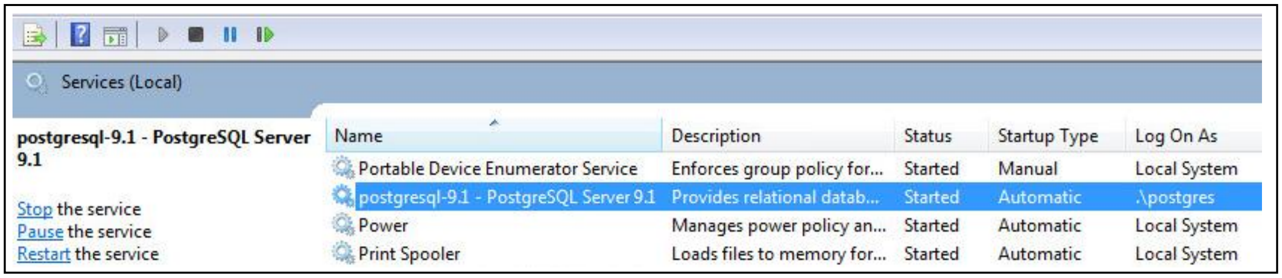
The three virtual machines (ERDP-NN, ERDP-01, and ERDP-Win7) were shipped with the ERW components pre-installed and ready for operation out-of-the-box. The user simply needed to initialize the ERW components after the servers were started.

[§] GATE's ANNIE System. http://gate.ac.uk/ie/annie.html (accessed June 2013).

These following steps we used are general guidelines, assuming the configurations have not been changed. A strict installation sequence is not required and can be reordered if desired. Users interested in ERW installation and configuration on a clean machine can find full ERW installation procedure starting in section 3. Otherwise, users can use the pre-configured and pre-installed virtual machines to test the ERW components.

Installation Sequence

1. Start PostgreSQL Service on ERDP-Win7 (figure 1).
    - Log in to ERDP-Win7 as the "arl" user/administrator.
    - Use Windows Services Manager to start the PostgreSQL Service.



postgresql-9.1 - PostgreSQL Server 9.1

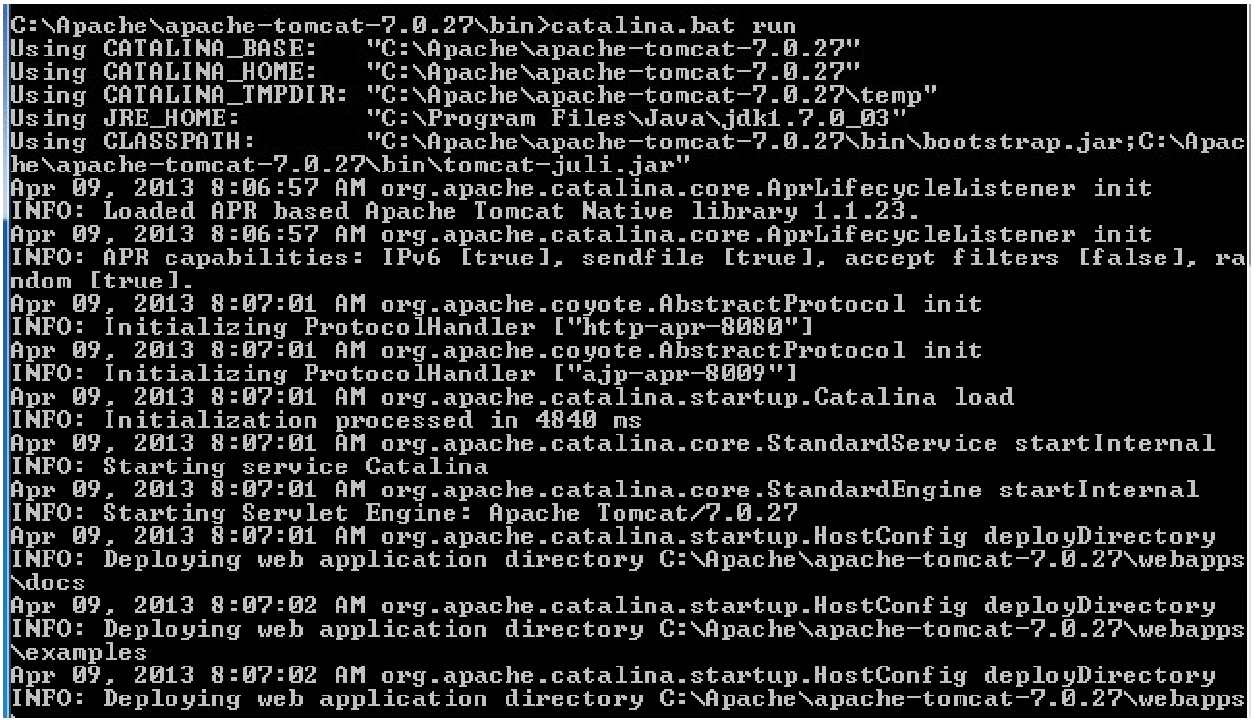
Stop the service
Pause the service
Restart the service

Name	Description	Status	Startup Type	Log On As
Portable Device Enumerator Service	Enforces group policy for...	Started	Manual	Local System
postgresql-9.1 - PostgreSQL Server 9.1	Provides relational datab...	Started	Automatic	.\postgres
Power	Manages power policy an...	Started	Automatic	Local System
Print Spooler	Loads files to memory for...	Started	Automatic	Local System

Figure 1. Starting PostgreSQL Service on Windows OS.

- The PostgreSQL Service is configured to start automatically and may have already started on its own after reboot.

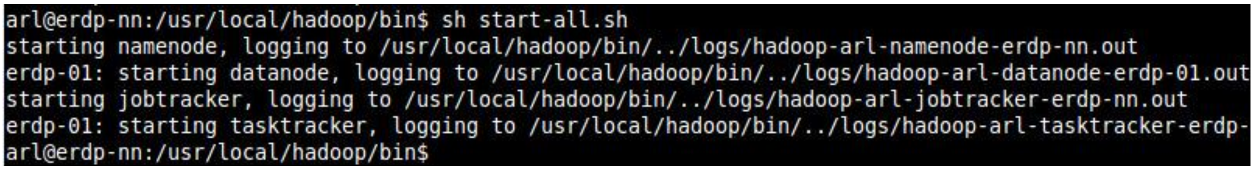
2. Start Global Graph REST Services on ERDP-Win7 (figure 2).
    - Open a new console window.
    - Change the directory to <CATALINA_HOME>\bin.
    - Run the following command: **catalina.bat run**



```
C:\Apache\apache-tomcat-7.0.27\bin>catalina.bat run
Using CATALINA_BASE:   "C:\Apache\apache-tomcat-7.0.27"
Using CATALINA_HOME:   "C:\Apache\apache-tomcat-7.0.27"
Using CATALINA_TMPDIR: "C:\Apache\apache-tomcat-7.0.27\temp"
Using JRE_HOME:        "C:\Program Files\Java\jdk1.7.0_03"
Using CLASSPATH:       "C:\Apache\apache-tomcat-7.0.27\bin\bootstrap.jar;C:\Apac
he\apache-tomcat-7.0.27\bin\tomcat-juli.jar"
Apr 09, 2013 8:06:57 AM org.apache.catalina.core.AprLifecycleListener init
INFO: Loaded APR based Apache Tomcat Native library 1.1.23.
Apr 09, 2013 8:06:57 AM org.apache.catalina.core.AprLifecycleListener init
INFO: APR capabilities: IPv6 [true], sendfile [true], accept filters [false], ra
ndom [true].
Apr 09, 2013 8:07:01 AM org.apache.coyote.AbstractProtocol init
INFO: Initializing ProtocolHandler ["http-apr-8080"]
Apr 09, 2013 8:07:01 AM org.apache.coyote.AbstractProtocol init
INFO: Initializing ProtocolHandler ["ajp-apr-8009"]
Apr 09, 2013 8:07:01 AM org.apache.catalina.startup.Catalina load
INFO: Initialization processed in 4840 ms
Apr 09, 2013 8:07:01 AM org.apache.catalina.core.StandardService startInternal
INFO: Starting service Catalina
Apr 09, 2013 8:07:01 AM org.apache.catalina.core.StandardEngine startInternal
INFO: Starting Servlet Engine: Apache Tomcat/7.0.27
Apr 09, 2013 8:07:01 AM org.apache.catalina.startup.HostConfig deployDirectory
INFO: Deploying web application directory C:\Apache\apache-tomcat-7.0.27\webapps
\docs
Apr 09, 2013 8:07:02 AM org.apache.catalina.startup.HostConfig deployDirectory
INFO: Deploying web application directory C:\Apache\apache-tomcat-7.0.27\webapps
\examples
Apr 09, 2013 8:07:02 AM org.apache.catalina.startup.HostConfig deployDirectory
INFO: Deploying web application directory C:\Apache\apache-tomcat-7.0.27\webapps
```

Figure 2. Starting Global Graph REST Service on Windows OS.

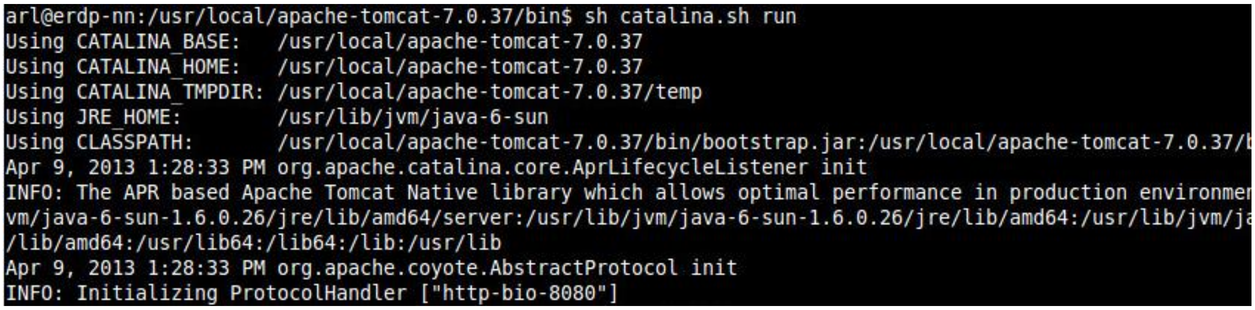
3. Start the Hadoop cluster on ERDP-NN (figure 3).
   - Log in to ERDP-NN as the “arl” user.
   - Open a new console window.
   - Change the directory to <HADOOP_HOME>/bin.
   - Run the following command: **sh start-all.sh**



```
arl@erdp-nn:/usr/local/hadoop/bin$ sh start-all.sh
starting namenode, logging to /usr/local/hadoop/bin/../logs/hadoop-arl-namenode-erdp-nn.out
erdp-01: starting datanode, logging to /usr/local/hadoop/bin/../logs/hadoop-arl-datanode-erdp-01.out
starting jobtracker, logging to /usr/local/hadoop/bin/../logs/hadoop-arl-jobtracker-erdp-nn.out
erdp-01: starting tasktracker, logging to /usr/local/hadoop/bin/../logs/hadoop-arl-tasktracker-erdp-
arl@erdp-nn:/usr/local/hadoop/bin$
```

Figure 3. Starting Hadoop cluster on Hadoop name node.

4. Start ERW Web Application on ERDP-NN (figure 4).
   - Open a new console window.
   - Change the directory to <CATALINA_HOME>/bin.
   - Run the following command: **catalina.sh run**



```
arl@erdp-nn:/usr/local/apache-tomcat-7.0.37/bin$ sh catalina.sh run
Using CATALINA_BASE:   /usr/local/apache-tomcat-7.0.37
Using CATALINA_HOME:   /usr/local/apache-tomcat-7.0.37
Using CATALINA_TMPDIR: /usr/local/apache-tomcat-7.0.37/temp
Using JRE_HOME:        /usr/lib/jvm/java-6-sun
Using CLASSPATH:       /usr/local/apache-tomcat-7.0.37/bin/bootstrap.jar:/usr/local/apache-tomcat-7.0.37/b
Apr 9, 2013 1:28:33 PM org.apache.catalina.core.AprLifecycleListener init
INFO: The APR based Apache Tomcat Native library which allows optimal performance in production environmer
vm/java-6-sun-1.6.0.26/jre/lib/amd64/server:/usr/lib/jvm/java-6-sun-1.6.0.26/jre/lib/amd64:/usr/lib/jvm/ja
/lib/amd64:/usr/lib64:/lib64:/lib:/usr/lib
Apr 9, 2013 1:28:33 PM org.apache.coyote.AbstractProtocol init
INFO: Initializing ProtocolHandler ["http-bio-8080"]
```

Figure 4. Starting the ERW Web application.

5. Open ERW Web application (figure 5).

    - Open the Internet browser to http://172.18.130.210:8080/EntityResolutionWebApp/. At the time of writing this report, the IP address of ERDP-NN was 172.18.130.210. The IP address on ERDP-NN will be different on a different setup. Please use the correct IP or equivalent domain name in the URL.

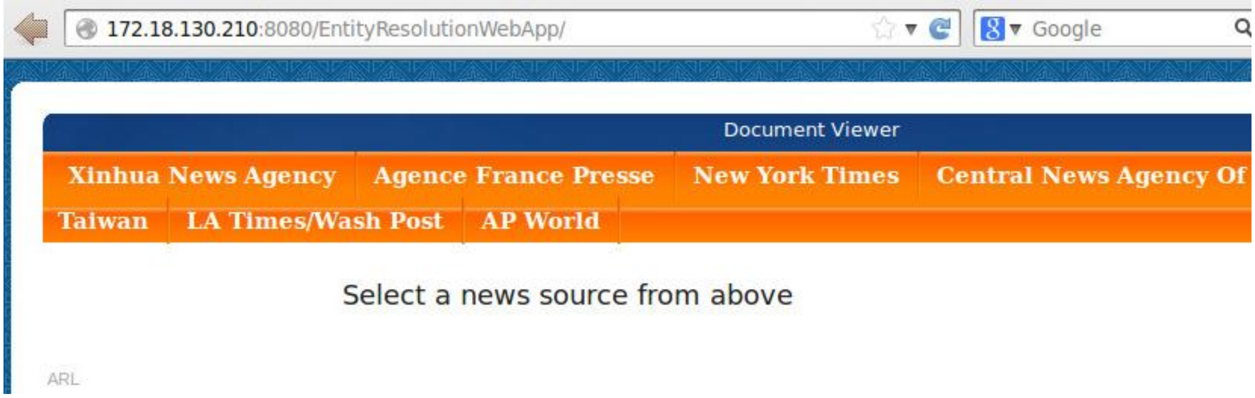

172.18.130.210:8080/EntityResolutionWebApp/
Google
Document Viewer
Xinhua News Agency
Agence France Presse
New York Times
Central News Agency Of Taiwan
LA Times/Wash Post
AP World
Select a news source from above
ARL

Figure 5. ERW Web application.

6. The ERW Web application is now ready for use. See section 6 for details on how to perform basic tasks.

3. Distribution

All the components (except Hadoop, Pig, and Apache Tomcat) needed and referenced in this report for ERW installation are included in “DISTRIBUTION_HOME”, which may be a DVD, flash drive, external hard disk, or zip archive. The string “<*DISTRIBUTION_HOME*>” is used in this report to refer to the “DISTRIBUTION_HOME” source.

4. Prerequisite Software

The following software should be installed on the target name node (“ERDP-NN”) according to their standard installation procedure:

- *Hadoop 0.20.1*

  Hadoop can be installed to any path, but this report refers to *HADOOP_HOME* as “/usr/local/hadoop-0.20.1” from this point on. Hadoop should also be configured and tested according to the user’s environment preference, and Hadoop-related environment variables should be set. For example, ERDP-NN’s Hadoop environment variables are the following:

  HADOOP_HOME=/usr/local/hadoop-0.20.1

  PATH=$PATH:/usr/local/hadoop-0.20.1/bin

  HADOOP_CONF_DIR=/usr/local/hadoop-0.20.1/conf

- *Pig 0.10.0*

  Pig can be installed to any path, but this report refers to PIG_HOME as “/usr/local/pig-0.10.0” from this point on. Pig should also be configured and tested according to the user’s environment preference, and Pig-related environment variables should be set. For example, ERDP-NN’s Pig environment variables are the following:

  PIG_HOME=/usr/local/pig-0.10.0

  PIGDIR=/usr/local/pig-0.10.0

  PIG_CLASSPATH=/usr/local/hadoop-0.20.1/conf

  PATH=$PATH:/usr/local/pig-0.10.0/bin

- *Apache Tomcat 7.0.37*

  Apache Tomcat can be installed to any path, but this report assumes the installed path is “/usr/local/apache-tomcat-7.0.37” from this point. Add the following environment variables:

  CATALINA_HOME=/usr/local/apache-tomcat-7.0.37

  PATH=$PATH:/usr/local/apache-tomcat-7.0.37/bin

- *Global Graph 1.4.6 and REST Services*

  Global Graph is the entity base of ERW and is a major process in the ERW installation process that requires installing additional components such as PostgreSQL, Apache Tomcat, Java, and PostGIS. The Global Graph 1.4.6 distribution files are under the "<DISTRIBUTION_HOME>/GlobalGraph/GG-1.4.6/globalgraph-dist-1.4.6-final/" directory. Potomac Fusion, Inc., (PFI) included a Global Graph installation document as part of the distribution, and is found in "<DISTRIBUTION_HOME>/GlobalGraph/globalgraph-dist-1.4.6-final/docs/Global_Graph_User_Guide.docx". In addition to the full installation document produced by PFI, an appendix "Installing Global Graph on ERDP-WIN7 (Windows 7 OS)" describing steps installing Global Graph for ERW is attached at the end of this report. Once Global Graph is successfully installed, Global Graph REST Web Service URL, user ID, and password will be needed for subsequent steps.

  Note: The user ID and password is stored in a plain text properties file so that ERW can query the Global Graph. Do not use security sensitive credentials for ERW.

5. Copy English Gigaword

English Gigaword (figure 6) is a 22.6-GB news archive that holds collections of news articles dated between 1994 and 2008 from six sources: Agence France Presse, AP World, Central News Agency of Taiwan, LA Times/Wash Post, New York Times, and Xinhua News Agency. English Gigaword is located in the *<DISTRIBUTION_HOME>/EnglishGigaword/* directory. There are six subdirectories (e.g., AFP_ENG, NYT_ENG, etc.), one for each news source. Each news source is listed below:

- *<DISTRIBUTION_HOME>/EnglishGigaword/AFP_ENG*

  Name of news source: Agence France Presse

  Monthly archived files: 122

  Size: 3.9 GB

- *<DISTRIBUTION_HOME>/EnglishGigaword/APW_ENG*

  Name of news source: AP World

  Monthly archived files: 170

  Size: 6.8 GB

- *<DISTRIBUTION_HOME>/EnglishGigaword/CNA_ENG*

  Name of news source: Central News Agency of Taiwan

  Monthly archived files: 120

  Size: 202 MB
- *<DISTRIBUTION_HOME>/EnglishGigaword/LTW_ENG*

  Name of news source: LA Times/Washington Post

  Monthly archived files: 173

  Size: 8.1 GB
- *<DISTRIBUTION_HOME>/EnglishGigaword/LTW_ENG*

  Name of news source: LA Times/Washington Post

  Monthly archived files: 114

  Size: 1.5 GB
- *<DISTRIBUTION_HOME>/EnglishGigaword/NYT_ENG*

  Name of news source: New York Times

  Monthly archived files: 173

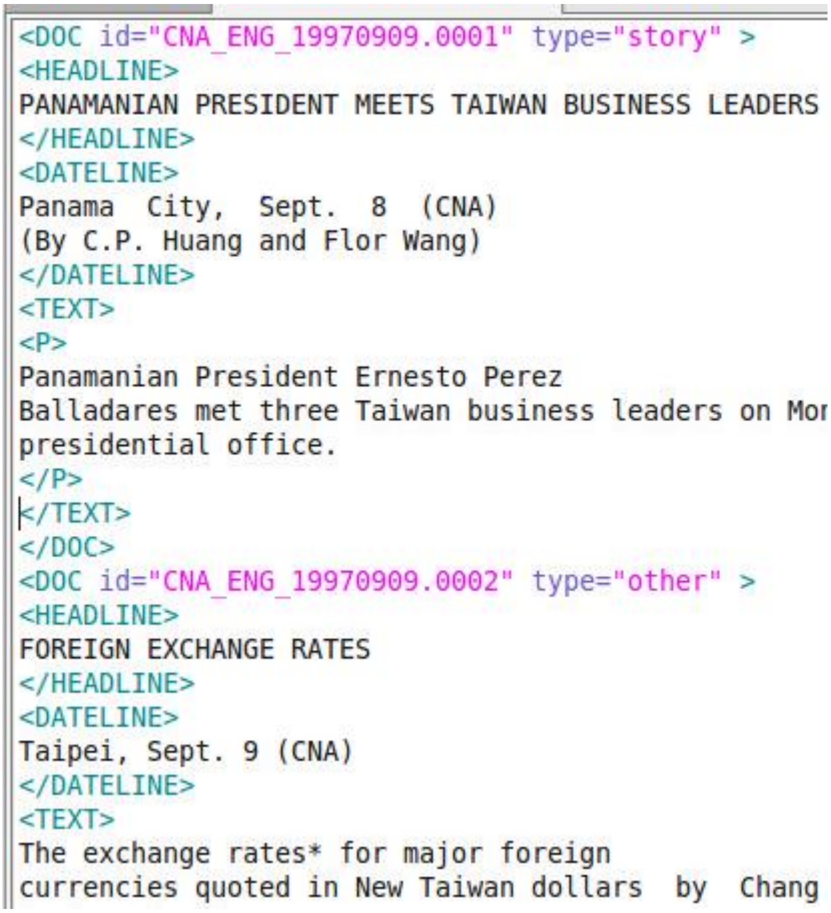
  Size: 8.1 GB
- *<DISTRIBUTION_HOME>/EnglishGigaword/XIN_ENG*

  Name of news source: Xinhua News Agency

  Monthly archived files: 167

  Size: 8.0 GB

Under each news source directory, there are files (ASCII encoded) that contain a concatenation of news articles published in a particular month. The number of articles in these files and the number files under each news source vary from each source. Each article is embedded with metadata specifying the document ID, headline, dateline, and text.



```
<DOC id="CNA_ENG_19970909.0001" type="story" >
<HEADLINE>
PANAMANIAN PRESIDENT MEETS TAIWAN BUSINESS LEADERS
</HEADLINE>
<DATELINE>
Panama  City,  Sept.  8  (CNA)
(By C.P. Huang and Flor Wang)
</DATELINE>
<TEXT>
<P>
Panamanian President Ernesto Perez
Balladares met three Taiwan business leaders on Mor
presidential office.
</P>
</TEXT>
</DOC>
<DOC id="CNA_ENG_19970909.0002" type="other" >
<HEADLINE>
FOREIGN EXCHANGE RATES
</HEADLINE>
<DATELINE>
Taipei, Sept. 9 (CNA)
</DATELINE>
<TEXT>
The exchange rates* for major foreign
currencies quoted in New Taiwan dollars  by  Chang
```

Figure 6. Sample English Gigaword monthly archived file.

Copy <*DISTRIBUTION_HOME*>*/EnglishGigaword/* to /data/EnglishGigaword/ on the name node and verify it is accessible by the “arl” user.

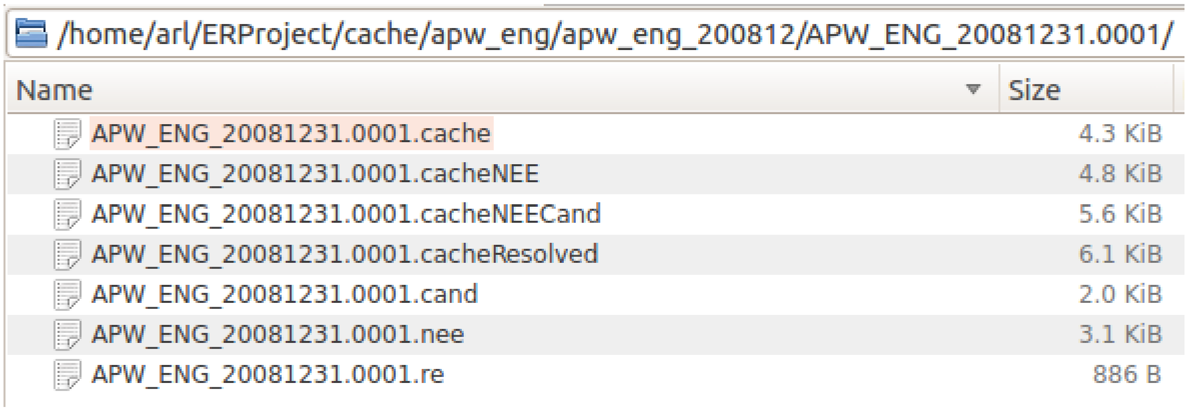
5.1 Install CUNY ENIE, Entity Base Cache, JARs, Pig Scripts, and Master Properties File

The majority of the ERW components are located in <*DISTRIBUTION_HOME*>*/ERProject* directory. Copy the entire <*DISTRIBUTION_HOME*>*/ERProject* directory to the user’s home directory (e.g., *~/ERProject),* which is referred to as <*ERProject*> from this point on. The ERProject directory is copied to the user’s home directory in order to reduce permission errors. The following notable subdirectories and files are found under ERProject directory.

- *Directory: “cache”*

  The “cache” directory is where temporary and intermediary files (e.g., copy of source file, named entity extraction (NEE) output, candidates output, hyperlinked file, etc.) for a selected document are stored in a subdirectory that preserves the hierarchy of the selected source article. For example, if a user selects a December 31, 2008, AP World news article

"Internet stocks fell in 2008 despite healthy signs," temporary files required for entity extraction, candidate identification, and entity resolution are saved to *<ERProject>/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001* (figure 7).



/home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/

Name	Size
APW_ENG_20081231.0001.cache	4.3 KiB
APW_ENG_20081231.0001.cacheNEE	4.8 KiB
APW_ENG_20081231.0001.cacheNEECand	5.6 KiB
APW_ENG_20081231.0001.cacheResolved	6.1 KiB
APW_ENG_20081231.0001.cand	2.0 KiB
APW_ENG_20081231.0001.nee	3.1 KiB
APW_ENG_20081231.0001.re	886 B

Figure 7. Intermediary files saved in the cache directory for an AP World article published on December 31, 2008.

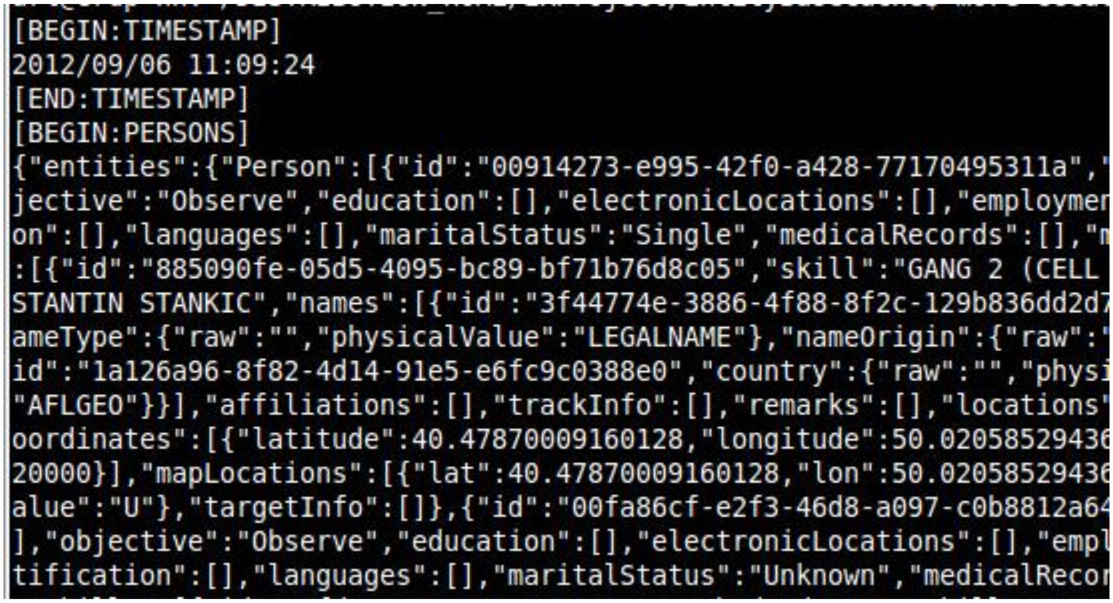
These intermediary files are cached so that the user will be able to instantly view results of old jobs. If these cache files are deleted, the user will have to re-run the process and wait for each step to complete before viewing the results. Deleting these cache files may be desirable if the source file has been altered and the user expects different results than the previous runs. In most cases, these temporary files can be saved indefinitely.

- *Directory: "CUNY"*

  The "CUNY" directory is where CUNY's ENIE is stored. ENIE is the entity extraction engine used in this configuration of ERW. Please contact CUNY for more details on the ENIE.

- *Directory: "EntityBaseCache"*

  The "EntityBaseCache" directory contains "GGCache.txt", which contains a dump (in JavaScript Object Notation [JSON] format) of all persons, organizations, and relationships from the Global Graph (figure 8). This local Global Graph cache is used for queries instead of the live Global Graph to reduce the time of processing. A single entity resolution process sends hundreds of relationship queries via REST Services, which has higher overhead than querying the local copy. If ERW detects that "GGCache.txt" does not exist, ERW will create a new "GGCahce.txt" and query the Global Graph for all persons, organizations, and relationships to save in the new file. Creating the new Entity Base Cache file will take several minutes (depending on how much data is in the Global Graph). The date and time the cache was created are saved on the first line of the "GGCache.txt".



```
[BEGIN:TIMESTAMP]
2012/09/06 11:09:24
[END:TIMESTAMP]
[BEGIN:PERSONS]
{"entities":{"Person":[{"id":"00914273-e995-42f0-a428-77170495311a","
jective":"Observe","education":[],"electronicLocations":[],"employmen
on":[],"languages":[],"maritalStatus":"Single","medicalRecords":[],"m
:[{"id":"885090fe-05d5-4095-bc89-bf71b76d8c05","skill":"GANG 2 (CELL
STANTIN STANKIC","names":[{"id":"3f44774e-3886-4f88-8f2c-129b836dd2d7
ameType":{"raw":"","physicalValue":"LEGALNAME"},"nameOrigin":{"raw":"
id":"1a126a96-8f82-4d14-91e5-e6fc9c0388e0","country":{"raw":"","physi
"AFLGEO"}}],"affiliations":[],"trackInfo":[],"remarks":[],"locations"
oordinates":[{"latitude":40.47870009160128,"longitude":50.02058529436
20000}],"mapLocations":[{"lat":40.47870009160128,"lon":50.02058529436
alue":"U"},"targetInfo":[]},{"id":"00fa86cf-e2f3-46d8-a097-c0b8812a64
],"objective":"Observe","education":[],"electronicLocations":[],"empl
tification":[],"languages":[],"maritalStatus":"Unknown","medicalRecor
```

Figure 8. Timestamp and person JSON in GGCache.txt.

Entity Base Caching can be disabled by entering an empty string for the “ENTITYBASE” property in the master properties file. If disabled, all Global Graph queries will be directed to the live Global Graph REST Service.

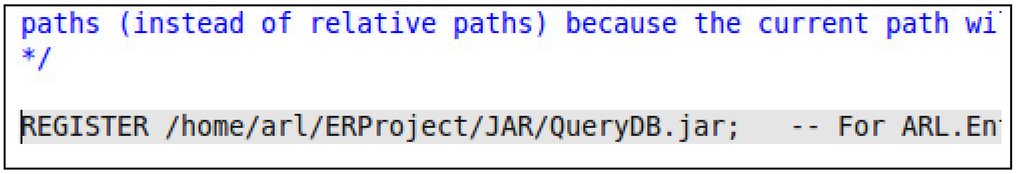
- *Directory: JAR*

  The “JAR” directory simply contains the Java archives (JARs) required by the ERW components.

- *Directory: PIGScripts*

  The “PIGScripts” directory contains Pig scripts used by ERW components to start jobs on the Hadoop cluster. Please visit the Hadoop Pig Web site (http://pig.apache.org/) for more details on PIG scripts.

  NOTE: After the PIG scripts are copied to the <ERProject>/PIGScripts directory, each script must be updated to correct the “REGISTER” command with correct path to dependent JAR(s) (figure 9). The “REGISTER” command will be the first non-comment line in the script, and the path will reference JAR(s) in the <ERProject>/JAR directory. The PIG script will immediately throw an error if the “REGISTER” path is not corrected.



```
paths (instead of relative paths) because the current path wi
*/

REGISTER /home/arl/ERProject/JAR/QueryDB.jar;   -- For ARL.En
```

Figure 9. “REGISTER” command in a PIG script.

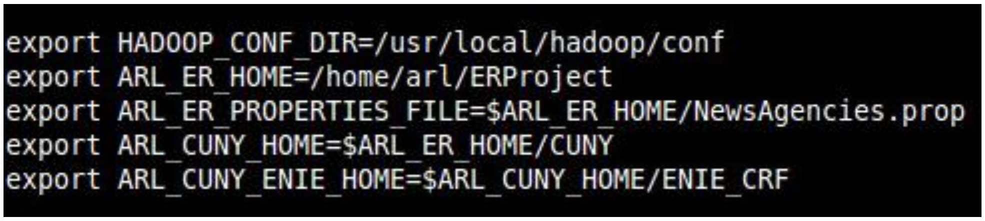
- *File: NewsAgencies.prop*

  “NewsAgencies.prop” file is the master properties file used by the ERW components. It contains properties required for Hadoop, Global Graph REST Service, REST Service client, CUNY ENIE, PIG scripts, the entity resolution algorithm, and English Gigaword. Setting this properties file is straightforward. Property names are self-explanatory and there are descriptions above most of the property names. This properties file must be closely reviewed for errors before starting ERW.

5.2 Configure Hadoop Data Node(s)

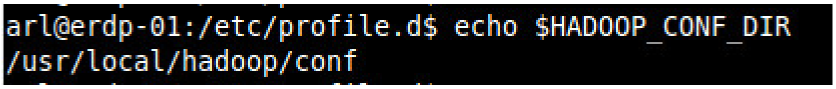
When a map-reduce job is initiated by a PIG script, that job is transferred to one of the data nodes that form the Hadoop cluster. The data nodes must be equipped and configured to process that job. This section describes the steps required to configure each data node. ERW Hadoop setup described in this report has only one data node (ERDP-01). For Hadoop clusters with many data nodes, these steps must be applied to each data node.

1. Add the following environment variables: HADOOP_CONF_DIR, ARL_ER_HOME, ARL_ER_PROPERTIES_FILE, ARL_CUNY_HOME, and ARL_CUNY_ENIE_HOME. On ERDP-01, these environment variables were defined in */etc/profile.d/arl_er.sh* (Setting environment variables will vary from machine to machine; please use the appropriate means of setting new environment variables on the data node). Root privilege is required to create this file. Change the permission of the file by running “*chmod 755 arl_er.sh*” (figure 10). Restart the server to activate the new environment variables, and verify the new variables are active by typing “*echo $HADOOP_CONF_DIR*” (figure 11).



```
export HADOOP_CONF_DIR=/usr/local/hadoop/conf
export ARL_ER_HOME=/home/arl/ERProject
export ARL_ER_PROPERTIES_FILE=$ARL_ER_HOME/NewsAgencies.prop
export ARL_CUNY_HOME=$ARL_ER_HOME/CUNY
export ARL_CUNY_ENIE_HOME=$ARL_CUNY_HOME/ENIE_CRF
```

Figure 10. Environment variables defined in /etc/profile.d/arl_er.sh.



```
arl@erdp-01:/etc/profile.d$ echo $HADOOP_CONF_DIR
/usr/local/hadoop/conf
```

Figure 11. Verify environment variables are active.

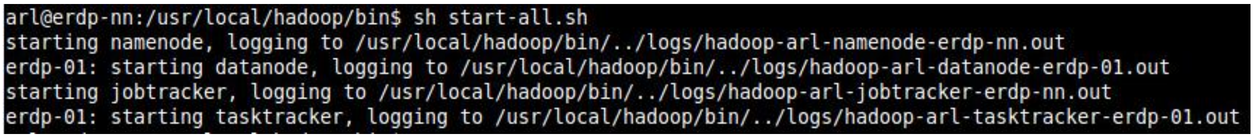
2. Create “ERProject” directory on the data node using the same <*ERProject*> path defined in ERDP-NN (i.e., path <*ERProject*> should be identical on ERDP-NN and ERDP-01). Copy the following directories and files from ERDP-NN to ERDP-01 using the same path as source path:

   - <ERProject>/cache
   - <ERProject>/CUNY
   - <ERProject>/EntityBaseCache
   - <ERProject>/NewsAgencies.prop

5.3 Test Hadoop Cluster, Hadoop File System (HDFS), and ERW Components via PIG Scripts

At this point, most ERW components are installed. These components must now be individually tested to verify they are configured correctly before continuing. There are three major steps in the ERW process: entity extraction, candidate identification, and entity resolution. Each step is tested by simulating the PIG script command built and executed by the ERW Web application during the ERW process. If a map-reduce PIG script test fails, it may be difficult to identify the problem because debug messages are only logged in the data nodes. Please refer to appendix B for tips on identifying map-reduce jobs.

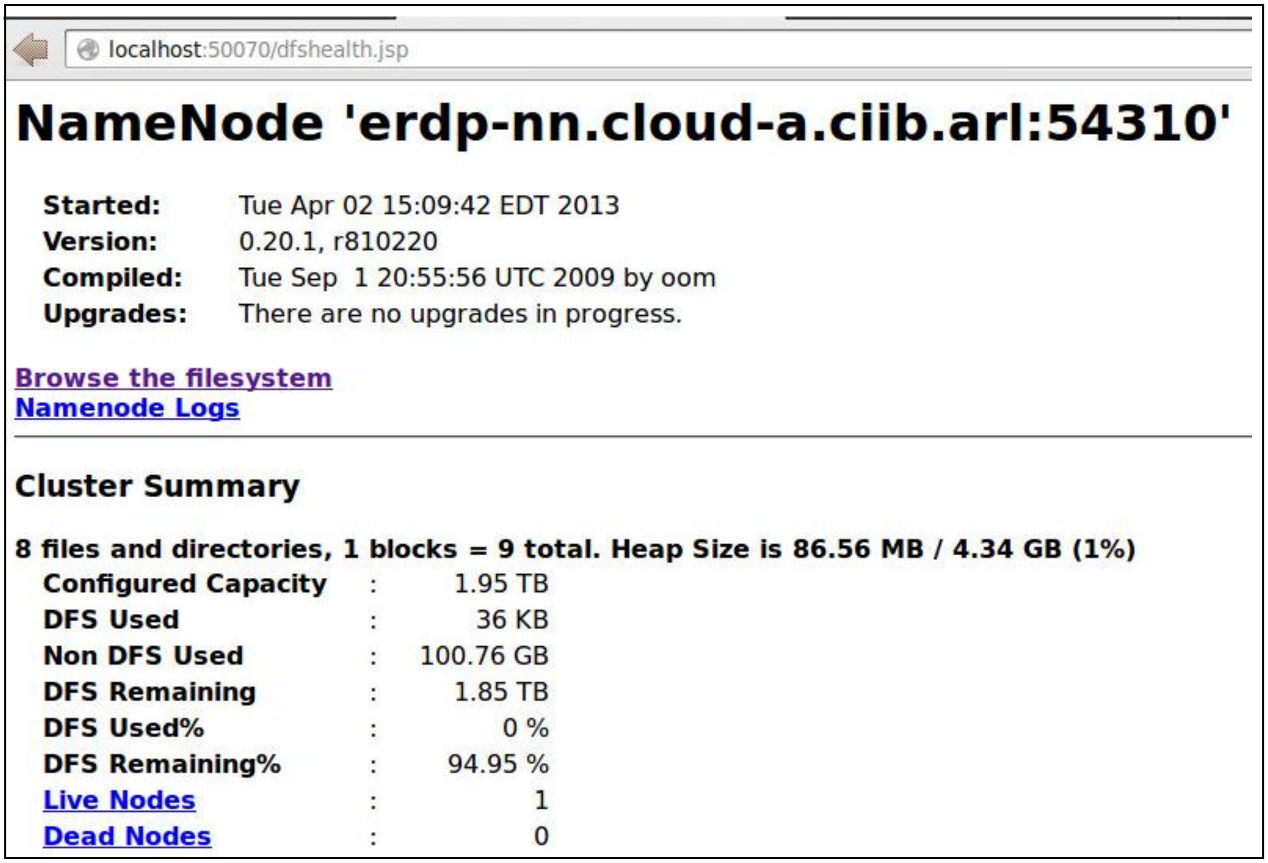
1. Before running these tests, start the Hadoop cluster by running the script “<HADOOP_HOME>/bin/start-all.sh” on ERDP-NN (figure 12).



```
arl@erdp-nn:/usr/local/hadoop/bin$ sh start-all.sh
starting namenode, logging to /usr/local/hadoop/bin/../logs/hadoop-arl-namenode-erdp-nn.out
erdp-01: starting datanode, logging to /usr/local/hadoop/bin/../logs/hadoop-arl-datanode-erdp-01.out
starting jobtracker, logging to /usr/local/hadoop/bin/../logs/hadoop-arl-jobtracker-erdp-nn.out
erdp-01: starting tasktracker, logging to /usr/local/hadoop/bin/../logs/hadoop-arl-tasktracker-erdp-01.out
```

Figure 12. Starting the Hadoop cluster.

2. Verify the Hadoop cluster and HDFS status page (figure 13) is accessible by browsing to http://localhost:50070/dfshealth.jsp and http://localhost:50070/nn_browsedfscontent.jsp (figure 14). The URL can be modified by replacing “localhost” with the ERDP-01 IP address or domain name to view it from another server.



localhost:50070/dfshealth.jsp

NameNode 'erdp-nn.cloud-a.ciib.arl:54310'

Started:	Tue Apr 02 15:09:42 EDT 2013
Version:	0.20.1, r810220
Compiled:	Tue Sep 1 20:55:56 UTC 2009 by oom
Upgrades:	There are no upgrades in progress.

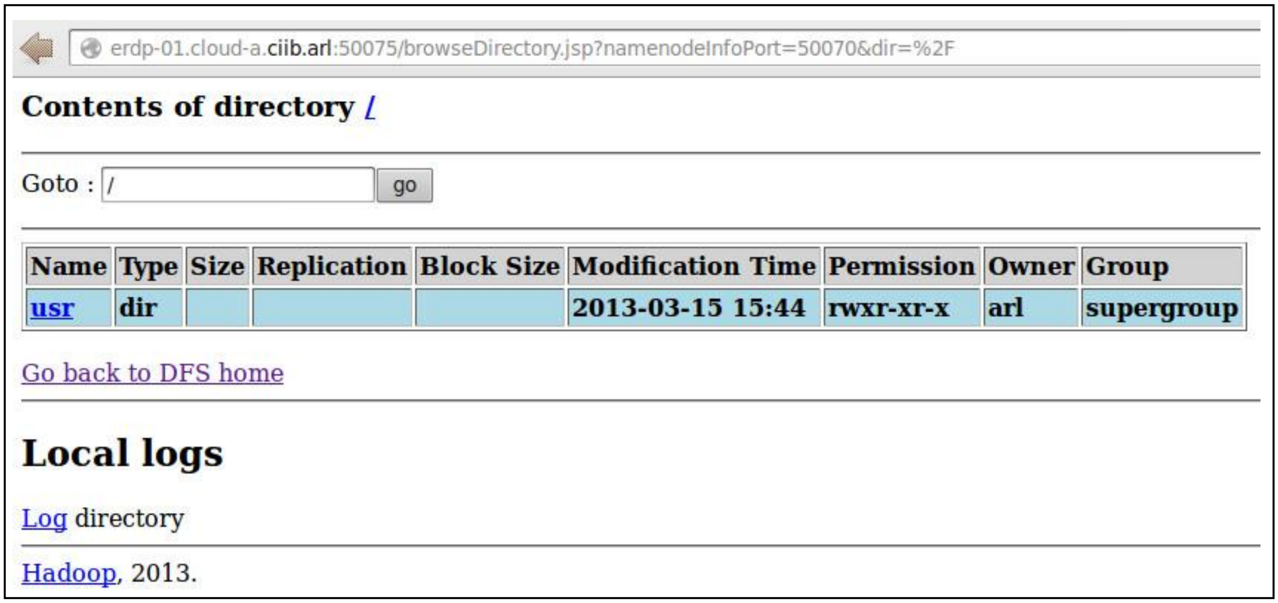
Browse the filesystem
Namenode Logs

Cluster Summary

8 files and directories, 1 blocks = 9 total. Heap Size is 86.56 MB / 4.34 GB (1%)

Configured Capacity	:	1.95 TB
DFS Used	:	36 KB
Non DFS Used	:	100.76 GB
DFS Remaining	:	1.85 TB
DFS Used%	:	0 %
DFS Remaining%	:	94.95 %
Live Nodes	:	1
Dead Nodes	:	0

Figure 13. Hadoop cluster status page.



erdp-01.cloud-a.ciib.arl:50075/browseDirectory.jsp?namenodeInfoPort=50070&dir=%2F

Contents of directory /

Goto : / go

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
usr	dir				2013-03-15 15:44	rwxr-xr-x	arl	supergroup

Go back to DFS home

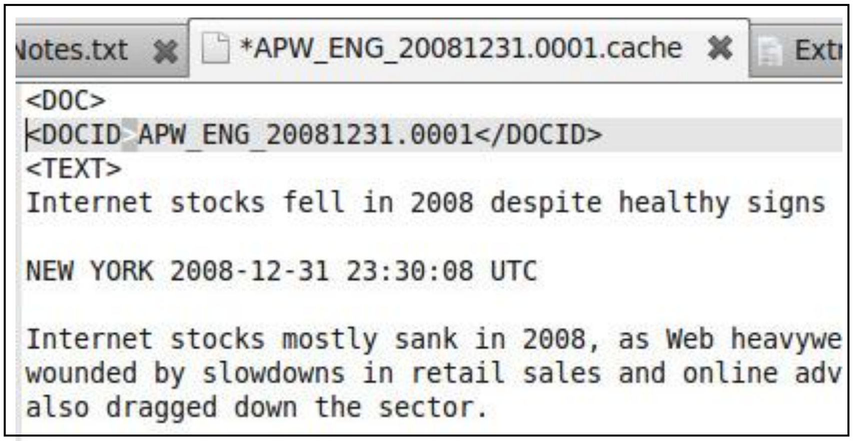
Local logs

Log directory

Hadoop, 2013.

Figure 14. HDFS browser.

3. *Copy <DISTRIBUTION_HOME>/TestFiles/APW_ENG_20081231.0001.cache to <ERProject>/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/ APW_ENG_20081231.0001.cache*. Figure 15 shows an AP World news article printed on December 31, 2008, extracted from the English Gigaword. This file will be used as a source file for testing.



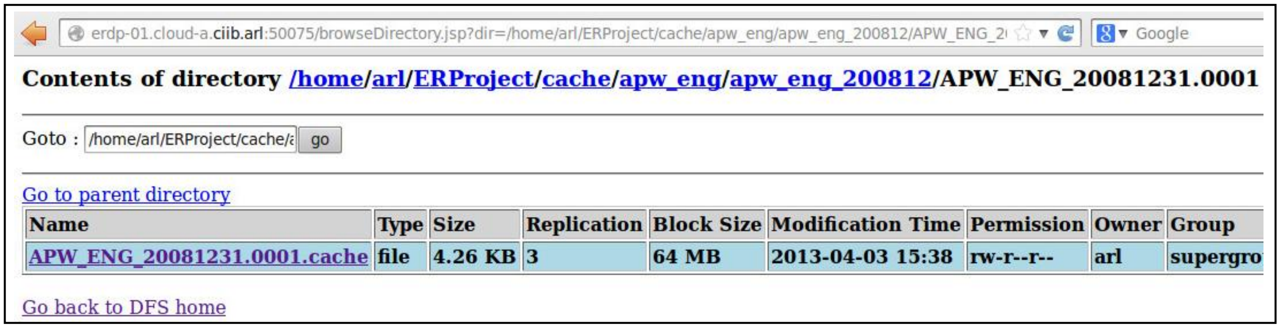
```
<DOC>
<DOCID APW_ENG_20081231.0001</DOCID>
<TEXT>
Internet stocks fell in 2008 despite healthy signs

NEW YORK 2008-12-31 23:30:08 UTC

Internet stocks mostly sank in 2008, as Web heavywe
wounded by slowdowns in retail sales and online adv
also dragged down the sector.
```

Figure 15. AP World article (APW_ENG_20081231.0001.cache) used for testing.

Copy <DISTRIBUTION_HOME>/TestFiles/APW_ENG_20081231.0001.cache to <ERProject>/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/ APW_ENG_20081231.0001.cache on HDFS (not the local file system as done in the previous step) (figure 16). Note that the target full paths are identical on the local file system and HDFS. The file can be copied to HDFS by running the following command:

```
hadoop fs -copyFromLocal
/home/arl/DISTRIBUTION_HOME/TestFiles/APW_ENG_20081231.0001.cache
/home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.00
01/APW_ENG_20081231.0001.cache
```

Verify that the test file is copied to HDFS by using the HDFS browser (figure 17).


erdp-01.cloud-a.ciib.arl:50075/browseDirectory.jsp?dir=/home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG_2(
Google

Contents of directory /home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001

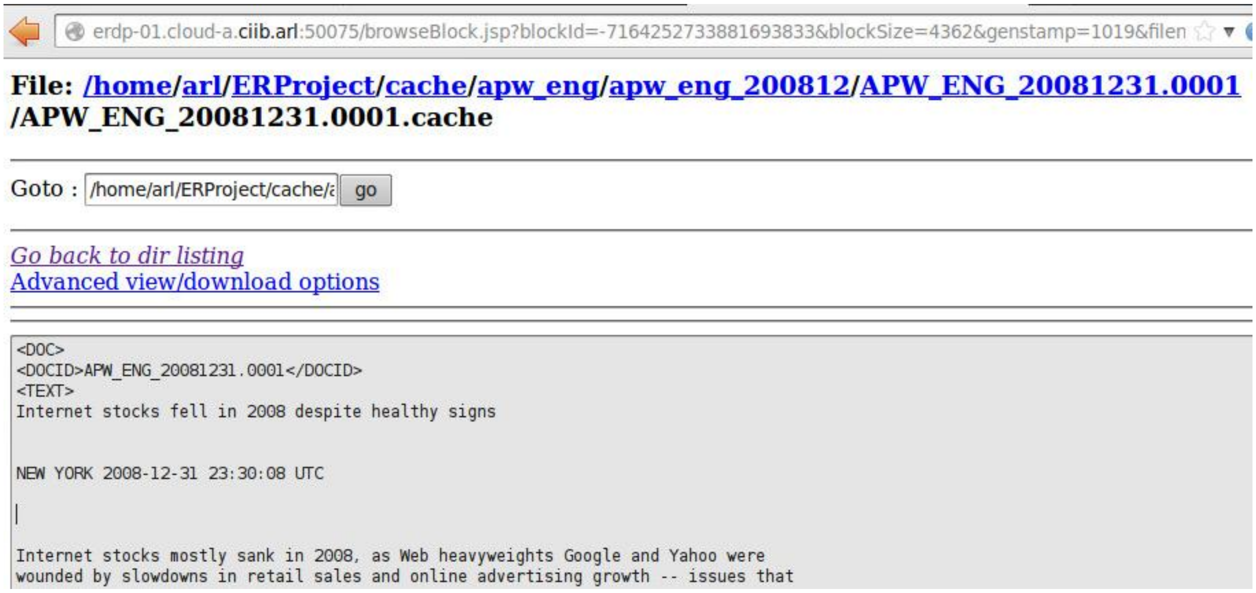
Goto : /home/arl/ERProject/cache/a go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
APW_ENG_20081231.0001.cache	file	4.26 KB	3	64 MB	2013-04-03 15:38	rw-r--r--	arl	supergro

Go back to DFS home

Figure 16. AP World article (APW_ENG_20081231.0001.cache) copied from local FS to HDFS.



erdp-01.cloud-a.ciib.arl:50075/browseBlock.jsp?blockId=-7164252733881693833&blockSize=4362&genstamp=1019&filen

File: /home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001
/APW_ENG_20081231.0001.cache

Goto : /home/arl/ERProject/cache/a go

Go back to dir listing
Advanced view/download options

```
<DOC>
<DOCID>APW_ENG_20081231.0001</DOCID>
<TEXT>
Internet stocks fell in 2008 despite healthy signs


NEW YORK 2008-12-31 23:30:08 UTC

|

Internet stocks mostly sank in 2008, as Web heavyweights Google and Yahoo were
wounded by slowdowns in retail sales and online advertising growth -- issues that
```

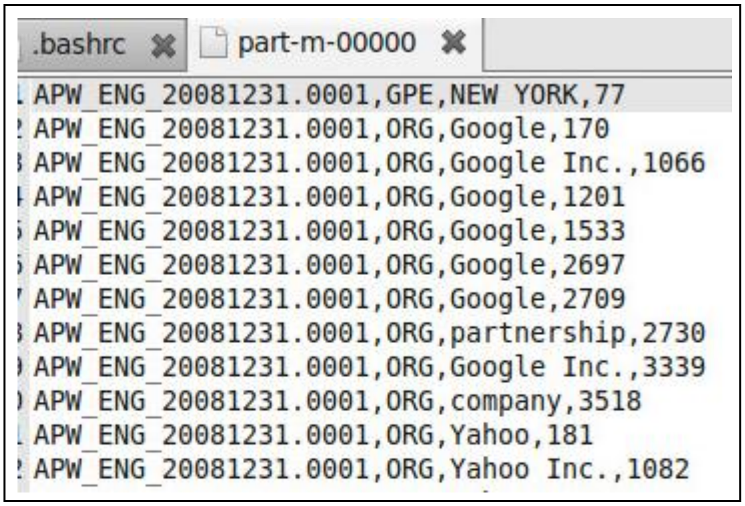
Figure 17. View of AP World article (APW_ENG_20081231.0001.cache) on HDFS.

4. Test the entity extraction Pig script in local mode with the following command:

```
java -cp /usr/local/pig-0.10.0/pig-0.10.0.jar org.apache.pig.Main -x local -param
outputNEDir="/home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG
_20081231.0001/APW_ENG_20081231.0001.nee" -param
inputDocFile="/home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG
_20081231.0001/APW_ENG_20081231.0001.cache" -param
pathPropFile="/home/arl/ERProject/NewsAgencies.prop"
/home/arl/ERProject/PIGScripts/ExtractEntitiesUsingCUNYENIECRF.pig
```

Once the PIG script completes, the entity extraction output is saved in *<ERProject>/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/APW_ENG_20081231.0001.nee/part-m-00000*. This output file (figure 18) is not a direct output from CUNY ENIE, it is a normalized summary generated by the entity extraction user defined function (UDF). Each line represents an entity identified by CUNY ENIE, and is formatted as follows:

[DOCID],[ENTITY_TYPE],[VALUE],[OFFSET]



```
APW_ENG_20081231.0001,GPE,NEW YORK,77
APW_ENG_20081231.0001,ORG,Google,170
APW_ENG_20081231.0001,ORG,Google Inc.,1066
APW_ENG_20081231.0001,ORG,Google,1201
APW_ENG_20081231.0001,ORG,Google,1533
APW_ENG_20081231.0001,ORG,Google,2697
APW_ENG_20081231.0001,ORG,Google,2709
APW_ENG_20081231.0001,ORG,partnership,2730
APW_ENG_20081231.0001,ORG,Google Inc.,3339
APW_ENG_20081231.0001,ORG,company,3518
APW_ENG_20081231.0001,ORG,Yahoo,181
APW_ENG_20081231.0001,ORG,Yahoo Inc.,1082
```

Figure 18. Output of entity extraction PIG script.

Delete the *<ERProject>/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/APW_ENG_20081231.0001.nee* directory before continuing to the next test.

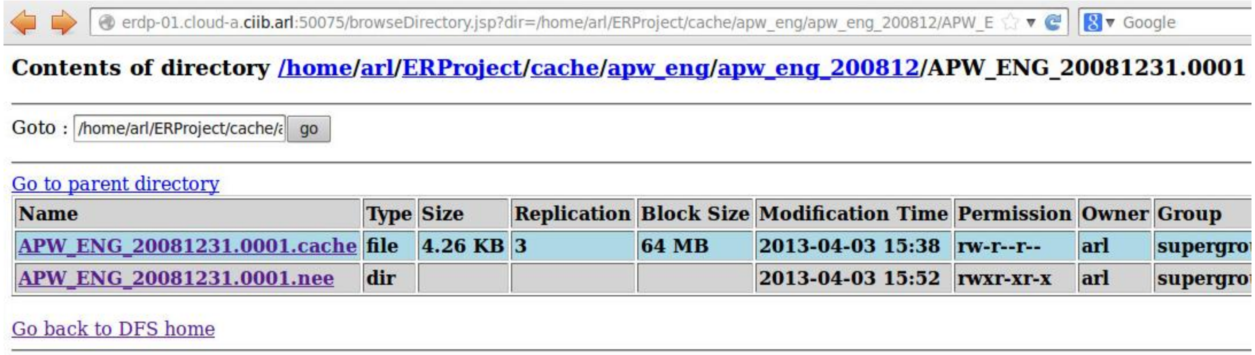
5. Test Entity Extraction Pig script in map-reduce mode with the following command:

```
java -Djavax.xml.parsers.DocumentBuilderFactory=com.sun.org.apache.xerces.internal.jaxp.DocumentBuilderFactoryImpl -cp /usr/local/hadoop/hadoop-0.20.1-core.jar:/usr/local/hadoop/lib/*:/usr/local/hadoop/conf:/usr/local/pig-0.10.0/pig-0.10.0-withouthadoop.jar org.apache.pig.Main -param outputNEDir="/home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/APW_ENG_20081231.0001.nee" -param inputDocFile="/home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/APW_ENG_20081231.0001.cache" -param pathPropFile="/home/arl/ERProject/NewsAgencies.prop" /home/arl/ERProject/PIGScripts/ExtractEntitiesUsingCUNYENIECRF.pig
```

Once the map-reduce PIG script completes, the entity extraction output (figure 19) is saved in the following:

<ERProject>/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/APW_ENG_20081231.0001.nee/part-m-00000 (*in HDFS*, not the local file system).

The content of this output file (figure 20) is identical to the output created in the previous test.



erdp-01.cloud-a.ciib.arl:50075/browseDirectory.jsp?dir=/home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_E

Contents of directory /home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001

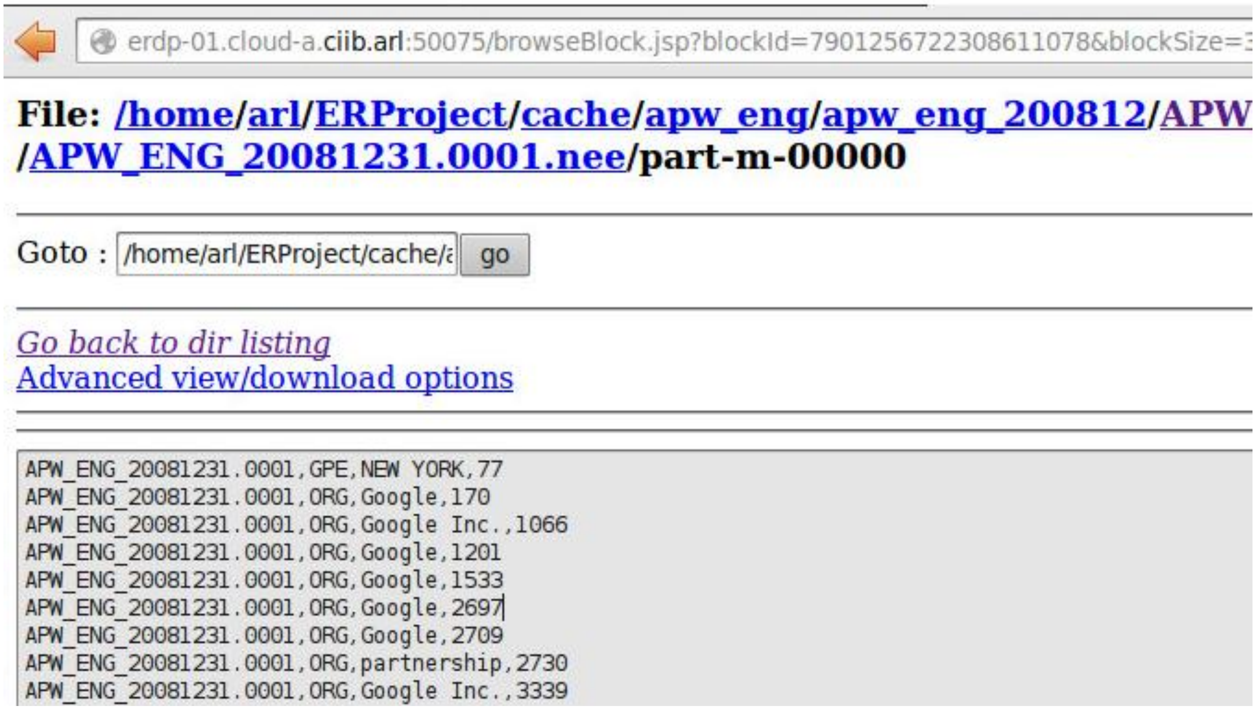
Goto : /home/arl/ERProject/cache/a go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
APW_ENG_20081231.0001.cache	file	4.26 KB	3	64 MB	2013-04-03 15:38	rw-r--r--	arl	supergro
APW_ENG_20081231.0001.nee	dir				2013-04-03 15:52	rwxr-xr-x	arl	supergro

Go back to DFS home

Figure 19. Entity extraction output on HDFS.



erdp-01.cloud-a.ciib.arl:50075/browseBlock.jsp?blockId=7901256722308611078&blockSize=3

File: /home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW /APW_ENG_20081231.0001.nee/part-m-00000

Goto : /home/arl/ERProject/cache/a go

Go back to dir listing
Advanced view/download options

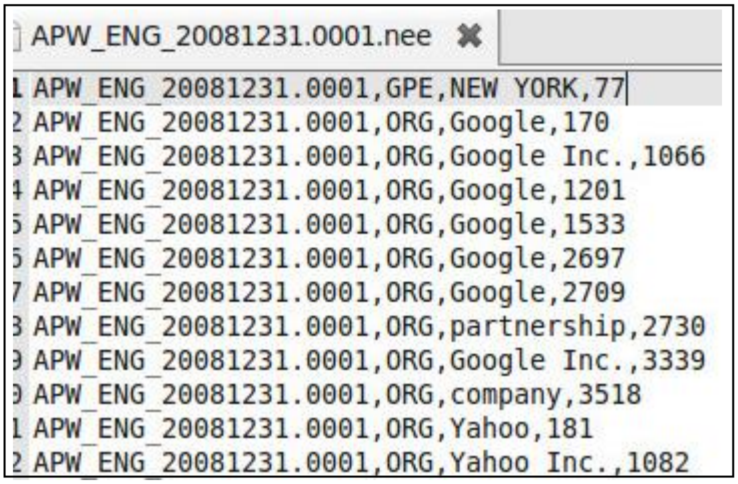
```
APW_ENG_20081231.0001,GPE,NEW YORK,77
APW_ENG_20081231.0001,ORG,Google,170
APW_ENG_20081231.0001,ORG,Google Inc.,1066
APW_ENG_20081231.0001,ORG,Google,1201
APW_ENG_20081231.0001,ORG,Google,1533
APW_ENG_20081231.0001,ORG,Google,2697
APW_ENG_20081231.0001,ORG,Google,2709
APW_ENG_20081231.0001,ORG,partnership,2730
APW_ENG_20081231.0001,ORG,Google Inc.,3339
```

Figure 20. View of the entity extraction output on HDFS.

Delete the *<ERProject>/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/APW_ENG_20081231.0001.nee* directory in HDFS before continuing to the next test. The directory can be deleted by running the following command:

hadoop fs -rmr /home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/APW_ENG_20081231.0001.nee

6. Copy <DISTRIBUTION_HOME>/TestFiles/APW_ENG_20081231.0001.nee to <ERProject>/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/ APW_ENG_20081231.0001.nee. Note that the “APW_ENG_80081231.0001.nee” referenced here is a *file*, not a directory as previously generated by the entity extraction Pig script from step 4 (figure 21). However, this file contains the same output entity list generated by the entity extraction Pig script, and is used as source for testing candidate identification PIG script.



```
APW_ENG_20081231.0001,GPE,NEW YORK,77
APW_ENG_20081231.0001,ORG,Google,170
APW_ENG_20081231.0001,ORG,Google Inc.,1066
APW_ENG_20081231.0001,ORG,Google,1201
APW_ENG_20081231.0001,ORG,Google,1533
APW_ENG_20081231.0001,ORG,Google,2697
APW_ENG_20081231.0001,ORG,Google,2709
APW_ENG_20081231.0001,ORG,partnership,2730
APW_ENG_20081231.0001,ORG,Google Inc.,3339
APW_ENG_20081231.0001,ORG,company,3518
APW_ENG_20081231.0001,ORG,Yahoo,181
APW_ENG_20081231.0001,ORG,Yahoo Inc.,1082
```

Figure 21. NEE file used to test the candidate identification PIG script.

Copy <DISTRIBUTION_HOME>/TestFiles/APW_ENG_20081231.0001.nee to <ERProject>/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/ APW_ENG_20081231.0001.nee *on HDFS* (not the local file system as done in the previous step). File can be copied to HDFS by running the following command:

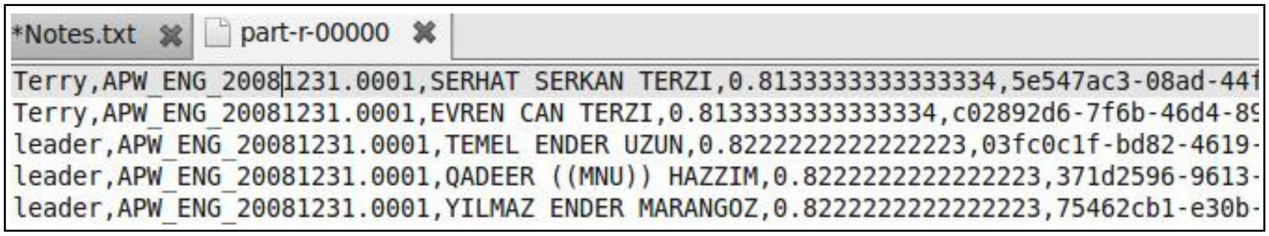
```
hadoop fs -copyFromLocal
/home/arl/DISTRIBUTION_HOME/TestFiles/APW_ENG_20081231.0001.nee
/home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001
/APW_ENG_20081231.0001.nee
```

7. Test candidate identification Pig script in local mode with the following command:

```
java -cp /usr/local/pig-0.10.0/pig-0.10.0.jar org.apache.pig.Main -x local -param
fileExtractedEntities="/home/arl/ERProject/cache/apw_eng/apw_eng_200812/AP
W_ENG_20081231.0001/APW_ENG_20081231.0001.nee" -param
pathEntitiesAndCandidatesOutput="/home/arl/ERProject/cache/apw_eng/apw_e
ng_200812/APW_ENG_20081231.0001/APW_ENG_20081231.0001.cand" -param
documentOfInterest="APW_ENG_20081231.0001" -param thresh="0.80" -
param pathEntityBase="/home/arl/ERProject/EntityBaseCache/GGCache.txt" -
param pathPropFile="/home/arl/ERProject/NewsAgencies.prop"
/home/arl/ERProject/PIGScripts/IdentifyCandidatesByGGWS.pig
```

Once the PIG script completes, candidate identification output is saved in *<ERProject>/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/APW_ENG_20081231.0001.cand/part-m-00000* (figure 22). Each line represents a candidate of an identified entity, and is formatted as follows:

```
[ENTITY_NAME],[DOCID],[CANDIDATE_NAME],[SS_SCORE],[CANDIDATE_
UUID]
```



```
Terry,APW_ENG_20081231.0001,SERHAT SERKAN TERZI,0.8133333333333334,5e547ac3-08ad-44f
Terry,APW_ENG_20081231.0001,EVREN CAN TERZI,0.8133333333333334,c02892d6-7f6b-46d4-8
leader,APW_ENG_20081231.0001,TEMEL ENDER UZUN,0.8222222222222223,03fc0c1f-bd82-4619-
leader,APW_ENG_20081231.0001,QADEER ((MNU)) HAZZIM,0.8222222222222223,371d2596-9613-
leader,APW_ENG_20081231.0001,YILMAZ ENDER MARANGOZ,0.8222222222222223,75462cb1-e30b-
```

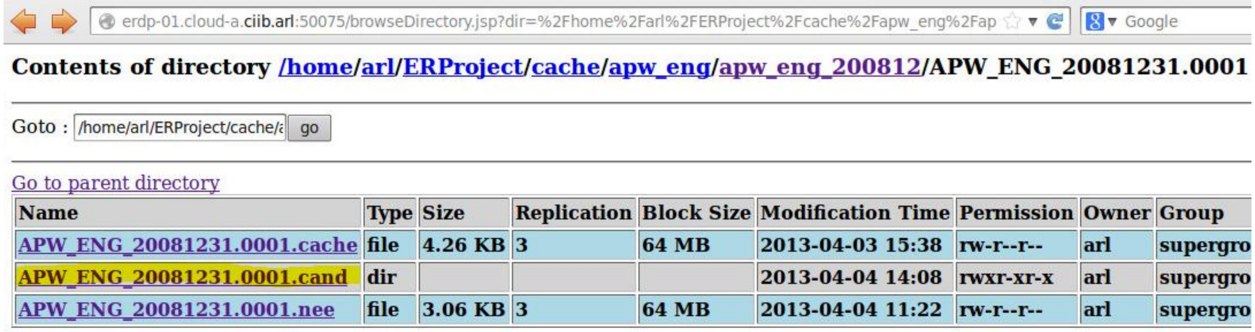
Figure 22. Output of candidate identification PIG script.

Delete the <ERProject>/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/APW_ENG_20081231.0001.cand directory before continuing to the next test.

8. Test the candidate identification Pig script in map-reduce mode with the following command:

```
java -
Djavax.xml.parsers.DocumentBuilderFactory=com.sun.org.apache.xerces.internal
.jaxp.DocumentBuilderFactoryImpl -cp /usr/local/hadoop/hadoop-0.20.1-
core.jar:/usr/local/hadoop/lib/*:/usr/local/hadoop/conf:/usr/local/pig-0.10.0/pig-
0.10.0.jar org.apache.pig.Main -param
fileExtractedEntities="/home/arl/ERProject/cache/apw_eng/apw_eng_200812/AP
W_ENG_20081231.0001/APW_ENG_20081231.0001.nee" -param
pathEntitiesAndCandidatesOutput="/home/arl/ERProject/cache/apw_eng/apw_e
ng_200812/APW_ENG_20081231.0001/APW_ENG_20081231.0001.cand" -param
documentOfInterest="APW_ENG_20081231.0001" -param thresh="0.80" -
param pathEntityBase="/home/arl/ERProject/EntityBaseCache/GGCache.txt" -
param pathPropFile="/home/arl/ERProject/NewsAgencies.prop"
/home/arl/ERProject/PIGScripts/IdentifyCandidatesByGGWS.pig
```

Once the map-reduce PIG script completes, the candidate identification output (figure 23) is saved in <ERProject>/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/APW_ENG_20081231.0001.cand/part-m-00000 (*in HDFS*, not the local file system). The content of this output file (figure 24) is identical to the output created in the previous test.


erdp-01.cloud-a.ciib.arl:50075/browseDirectory.jsp?dir=%2Fhome%2Farl%2FERProject%2Fcache%2Fapw_eng%2Fap
Google

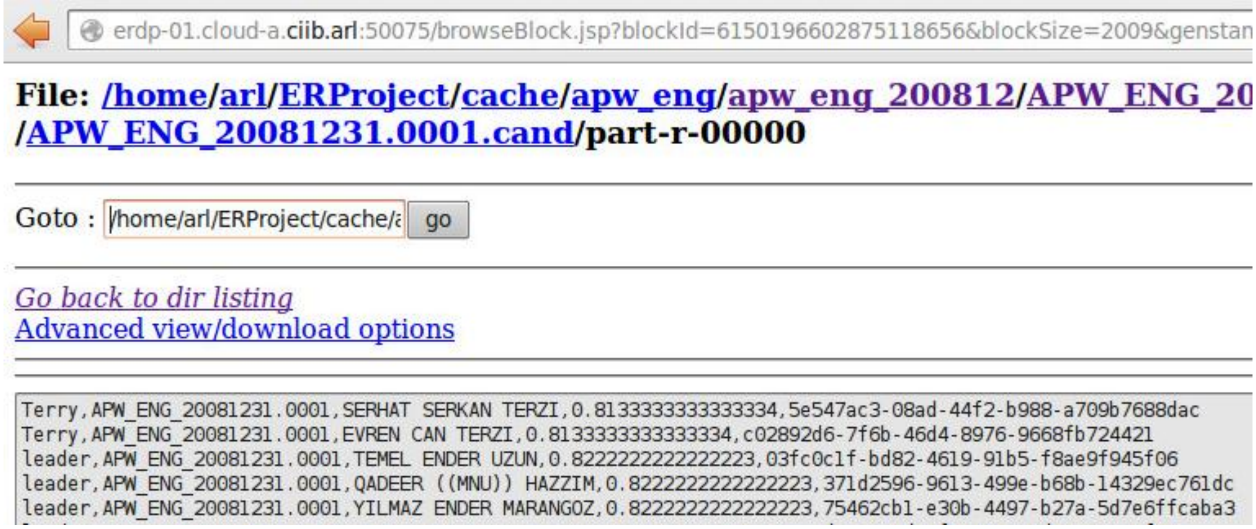
Contents of directory /home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001

Goto : /home/arl/ERProject/cache/a go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
APW_ENG_20081231.0001.cache	file	4.26 KB	3	64 MB	2013-04-03 15:38	rw-r--r--	arl	supergro
APW_ENG_20081231.0001.cand	dir				2013-04-04 14:08	rwxr-xr-x	arl	supergro
APW_ENG_20081231.0001.nee	file	3.06 KB	3	64 MB	2013-04-04 11:22	rw-r--r--	arl	supergro

Figure 23. Candidate identification output on HDFS.


erdp-01.cloud-a.ciib.arl:50075/browseBlock.jsp?blockId=6150196602875118656&blockSize=2009&genstan

File: /home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG_20
/APW_ENG_20081231.0001.cand/part-r-00000

Goto : /home/arl/ERProject/cache/a go

Go back to dir listing
Advanced view/download options

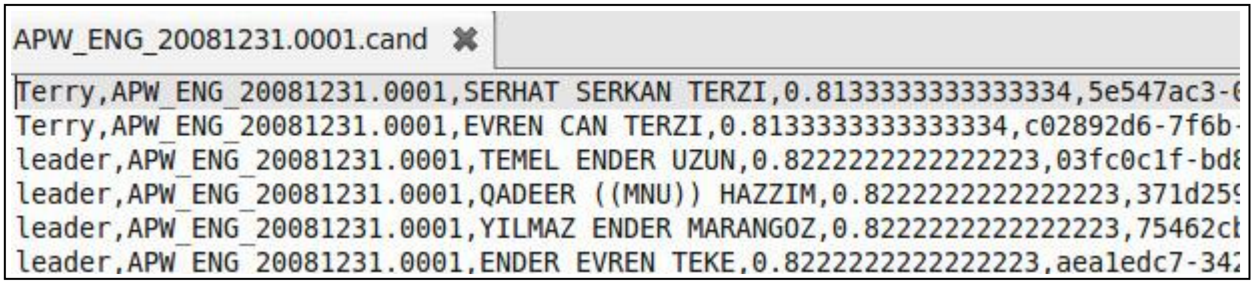
```
Terry,APW_ENG_20081231.0001,SERHAT SERKAN TERZI,0.8133333333333334,5e547ac3-08ad-44f2-b988-a709b7688dac
Terry,APW_ENG_20081231.0001,EVREN CAN TERZI,0.8133333333333334,c02892d6-7f6b-46d4-8976-9668fb724421
leader,APW_ENG_20081231.0001,TEMEL ENDER UZUN,0.8222222222222223,03fc0c1f-bd82-4619-91b5-f8ae9f945f06
leader,APW_ENG_20081231.0001,QADEER ((MNU)) HAZZIM,0.8222222222222223,371d2596-9613-499e-b68b-14329ec761dc
leader,APW_ENG_20081231.0001,YILMAZ ENDER MARANGOZ,0.8222222222222223,75462cb1-e30b-4497-b27a-5d7e6ffcaba3
```

Figure 24. View of the candidate identification output on HDFS.

Delete the *<ERProject>/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/APW_ENG_20081231.0001.cand* directory in *HDFS* before continuing to the next test. The directory can be deleted by running the following command:

hadoop fs -rmr /home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001 /APW_ENG_20081231.0001.cand

9. Copy *<DISTRIBUTION_HOME>/TestFiles/APW_ENG_20081231.0001.cand* to *<ERProject>/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/ APW_ENG_20081231.0001.cand*. Note that the "APW_ENG_80081231.0001.cand" referenced here is a *file*, not a directory as previously generated by the candidate identification Pig script from step 7. However, this file contains the same output candidate list generated by the candidate identification Pig script, and is used as source for testing entity resolution PIG script (figure 25).



```
APW_ENG_20081231.0001.cand
Terry,APW_ENG_20081231.0001,SERHAT SERKAN TERZI,0.8133333333333334,5e547ac3-
Terry,APW_ENG_20081231.0001,EVREN CAN TERZI,0.8133333333333334,c02892d6-7f6b-
leader,APW_ENG_20081231.0001,TEMEL ENDER UZUN,0.8222222222222223,03fc0c1f-bd
leader,APW_ENG_20081231.0001,QADEER ((MNU)) HAZZIM,0.8222222222222223,371d25
leader,APW_ENG_20081231.0001,YILMAZ ENDER MARANGOZ,0.8222222222222223,75462c
leader,APW_ENG_20081231.0001,ENDER EVREN TEKE,0.8222222222222223,aea1edc7-34
```

Figure 25. CAND file used to test the entity resolution PIG script.

Copy <DISTRIBUTION_HOME>/TestFiles/APW_ENG_20081231.0001.cand to <ERProject>/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/ APW_ENG_20081231.0001.cand *on HDFS* (not the local file system as done in the previous step). The file can be copied to HDFS by running the following command:

```
hadoop fs -copyFromLocal
/home/arl/DISTRIBUTION_HOME/TestFiles/APW_ENG_20081231.0001.cand
/home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001
/APW_ENG_20081231.0001.cand
```

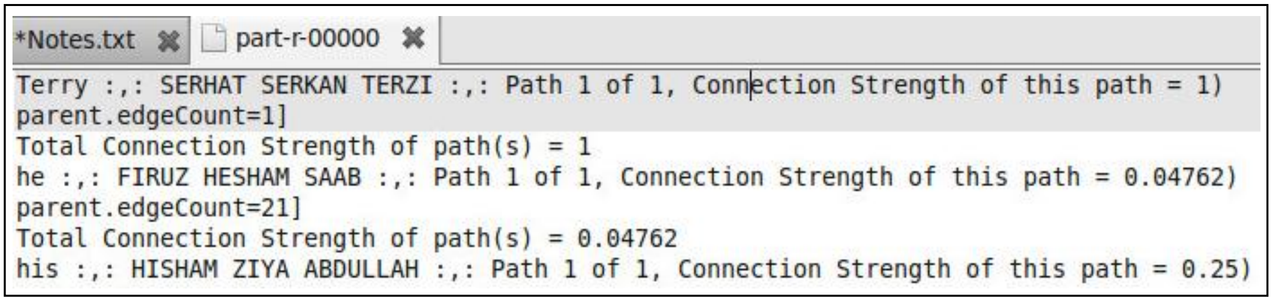
10. Test the entity resolution Pig script in local mode with the following command:

```
java -
Djavax.xml.parsers.DocumentBuilderFactory=com.sun.org.apache.xerces.internal
.jaxp.DocumentBuilderFactoryImpl -cp /usr/local/pig-0.10.0/pig-0.10.0.jar
org.apache.pig.Main -x local -param
documentOfInterest="APW_ENG_20081231.0001" -param
pathEntitiesAndCandidates="/home/arl/ERProject/cache/apw_eng/apw_eng_2008
12/APW_ENG_20081231.0001/APW_ENG_20081231.0001.cand" -param
fileResolvedEntitiesOutput="/home/arl/ERProject/cache/apw_eng/apw_eng_2008
12/APW_ENG_20081231.0001/APW_ENG_20081231.0001.re" -param
pathStorageLimit="25" -param maxPathLength="5" -param kRadius="4" -
param debugLevel="0" -param useGreedy="true" -param
pathEntityBase="/home/arl/ERProject/EntityBaseCache/GGCache.txt" -param
```

pathPropFile="/home/arl/ERProject/NewsAgencies.prop" /home/arl/ERProject/PIGScripts/EntityResolution_ManySrcToOneTrg.pig

Once the PIG script completes, the entity resolution output (figure 26) is saved in *<ERProject>/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/APW_ENG_20081231.0001.re/part-m-00000*. Each line represents a resolved entity, and is formatted:

[ENTITY_NAME] :,: [RESOLVED_ENTITY] :,: [PATH DETAILS]



```
Terry :,: SERHAT SERKAN TERZI :,: Path 1 of 1, Connection Strength of this path = 1)
parent.edgeCount=1]
Total Connection Strength of path(s) = 1
he :,: FIRUZ HESHAM SAAB :,: Path 1 of 1, Connection Strength of this path = 0.04762)
parent.edgeCount=21]
Total Connection Strength of path(s) = 0.04762
his :,: HISHAM ZIYA ABDULLAH :,: Path 1 of 1, Connection Strength of this path = 0.25)
```

Figure 26. Output of the entity resolution PIG script.

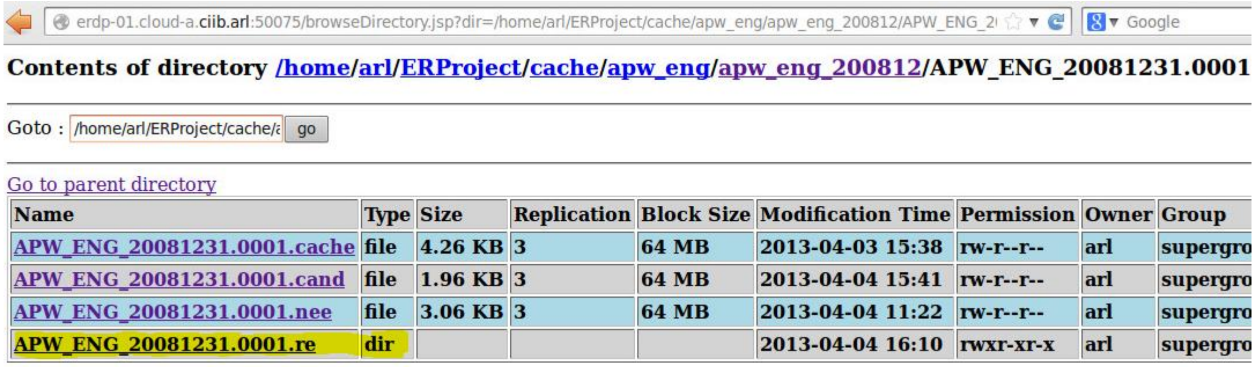
Delete the <ERProject>/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/APW_ENG_20081231.0001.re directory.

11. Test entity resolution Pig script in map-reduce mode with the following command:

java -Djavax.xml.parsers.DocumentBuilderFactory=com.sun.org.apache.xerces.internal.jaxp.DocumentBuilderFactoryImpl -cp /usr/local/hadoop/hadoop-0.20.1-core.jar:/usr/local/hadoop/lib/*:/usr/local/hadoop/conf:/usr/local/pig-0.10.0/pig-0.10.0.jar org.apache.pig.Main -param documentOfInterest="APW_ENG_20081231.0001" -param pathEntitiesAndCandidates="/home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/APW_ENG_20081231.0001.cand" -param fileResolvedEntitiesOutput="/home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/APW_ENG_20081231.0001.re" -param pathStorageLimit="25" -param maxPathLength="5" -param kRadius="4" -param debugLevel="0" -param useGreedy="true" -param pathEntityBase="/home/arl/ERProject/EntityBaseCache/GGCache.txt" -param pathPropFile="/home/arl/ERProject/NewsAgencies.prop" /home/arl/ERProject/PIGScripts/EntityResolution_ManySrcToOneTrg.pig

Once the map-reduce PIG script completes, entity resolution output (figure 27) will be saved in

<ERProject>/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/APW_ENG_20081231.0001.re/part-m-00000 (*in HDFS*, not the local file system) (figure 28). The content of this output file is identical to the output created in the previous test.



erdp-01.cloud-a.ciib.arl:50075/browseDirectory.jsp?dir=/home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG_2

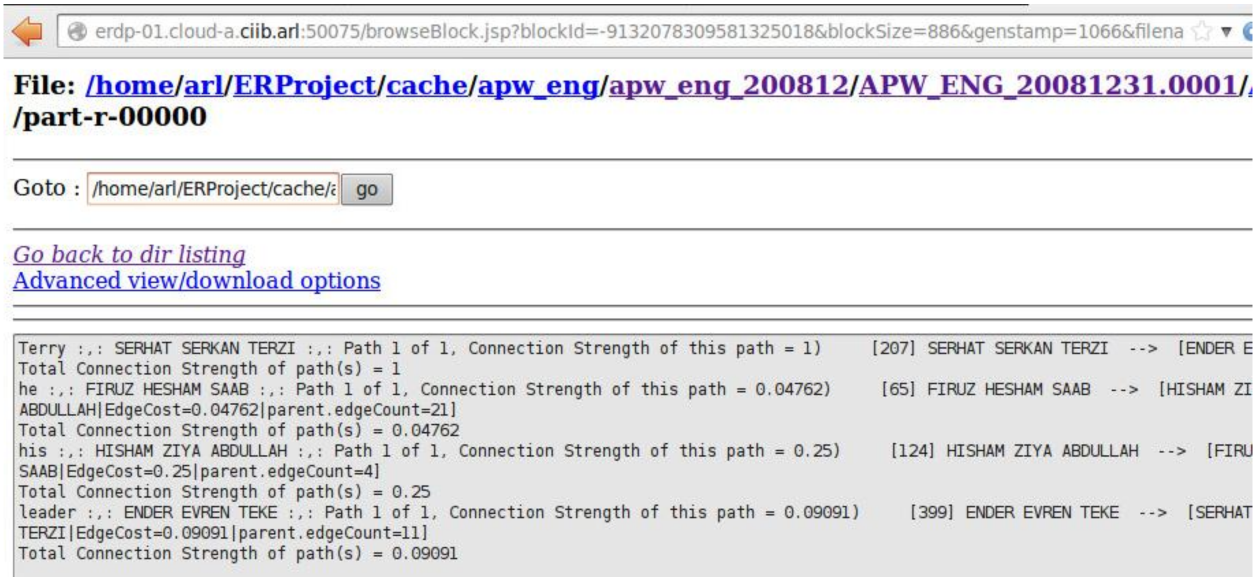
Contents of directory /home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001

Goto : /home/arl/ERProject/cache/a go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
APW_ENG_20081231.0001.cache	file	4.26 KB	3	64 MB	2013-04-03 15:38	rw-r--r--	arl	supergro
APW_ENG_20081231.0001.cand	file	1.96 KB	3	64 MB	2013-04-04 15:41	rw-r--r--	arl	supergro
APW_ENG_20081231.0001.nee	file	3.06 KB	3	64 MB	2013-04-04 11:22	rw-r--r--	arl	supergro
APW_ENG_20081231.0001.re	dir				2013-04-04 16:10	rwxr-xr-x	arl	supergro

Figure 27. Resolved entity output on HDFS.



erdp-01.cloud-a.ciib.arl:50075/browseBlock.jsp?blockId=-9132078309581325018&blockSize=886&genstamp=1066&filena

File: /home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/
/part-r-00000

Goto : /home/arl/ERProject/cache/a go

Go back to dir listing
Advanced view/download options

```
Terry :,: SERHAT SERKAN TERZI :,: Path 1 of 1, Connection Strength of this path = 1)     [207] SERHAT SERKAN TERZI  -->  [ENDER E
Total Connection Strength of path(s) = 1
he :,: FIRUZ HESHAM SAAB :,: Path 1 of 1, Connection Strength of this path = 0.04762)     [65] FIRUZ HESHAM SAAB  -->  [HISHAM ZI
ABDULLAH|EdgeCost=0.04762|parent.edgeCount=21]
Total Connection Strength of path(s) = 0.04762
his :,: HISHAM ZIYA ABDULLAH :,: Path 1 of 1, Connection Strength of this path = 0.25)     [124] HISHAM ZIYA ABDULLAH  -->  [FIRU
SAAB|EdgeCost=0.25|parent.edgeCount=4]
Total Connection Strength of path(s) = 0.25
leader :,: ENDER EVREN TEKE :,: Path 1 of 1, Connection Strength of this path = 0.09091)     [399] ENDER EVREN TEKE  -->  [SERHAT
TERZI|EdgeCost=0.09091|parent.edgeCount=11]
Total Connection Strength of path(s) = 0.09091
```

Figure 28. View of resolved entity output on HDFS.

Delete the *<ERProject>/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/APW_ENG_20081231.0001.re* directory in *HDFS*. The directory can be deleted by running the following command:

hadoop fs -rmr /home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/APW_ENG_20081231.0001.re

12. Copy <DISTRIBUTION_HOME>/TestFiles/APW_ENG_20081231.0001.re to <ERProject>/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/ APW_ENG_20081231.0001.re.

    Copy <DISTRIBUTION_HOME>/TestFiles/APW_ENG_20081231.0001.re to <ERProject>/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/ APW_ENG_20081231.0001.re **on HDFS.**

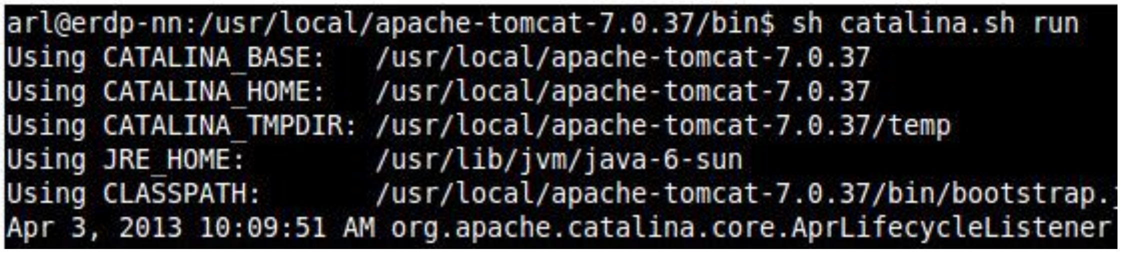
Testing is now complete. If all PIG scripts completed without any problems, then ERW components are installed and configured correctly. Section 5.4 describes installing the ERW front-end graphical user interface (GUI).

5.4 Deploy ERW Web Application

A Java Server Page (JSP) application was created to serve as an easy-to-use front-end GUI to the ERW and demonstrate each components of the ERW. The JSP application allows the user to select particular news article, run entity extraction, run candidate identification, run entity resolution, and view details of each resolved entity. Results of each step are automatically displayed to show what each component is contributing to the overall ERW process. In addition, complexities of entity extraction, candidate identification, entity resolution, and map-reduce jobs are hidden behind the scene and the user simply needs to click on links.

Follow these steps to install the ERW Web application:

1. Start Apache Tomcat by running "*<CATALINA_HOME>/bin/catalina.sh run*" (figure 29). There are several ways to start Tomcat, but this startup method is preferred because it keeps the console screen open while Tomcat is running and displays real-time status and debug messages from the ERW Web application. Verify by Tomcat is running by browsing to http://localhost:8080.

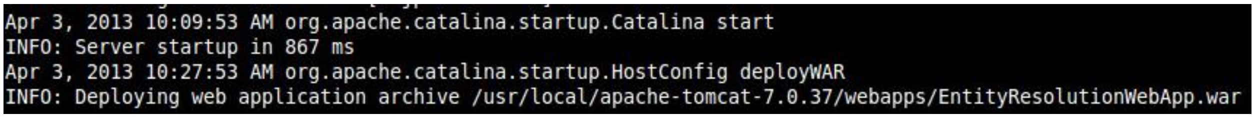


```
arl@erdp-nn:/usr/local/apache-tomcat-7.0.37/bin$ sh catalina.sh run
Using CATALINA_BASE:   /usr/local/apache-tomcat-7.0.37
Using CATALINA_HOME:   /usr/local/apache-tomcat-7.0.37
Using CATALINA_TMPDIR: /usr/local/apache-tomcat-7.0.37/temp
Using JRE_HOME:        /usr/lib/jvm/java-6-sun
Using CLASSPATH:       /usr/local/apache-tomcat-7.0.37/bin/bootstrap.
Apr 3, 2013 10:09:51 AM org.apache.catalina.core.AprLifecycleListener
```

Figure 29. Starting Apache Tomcat.

2. Copy the <DISTRIBUTION_HOME>/EntityResolutionWebApp.war to <CATALINA_HOME>/webapps directory. In few seconds, Tomcat will automatically

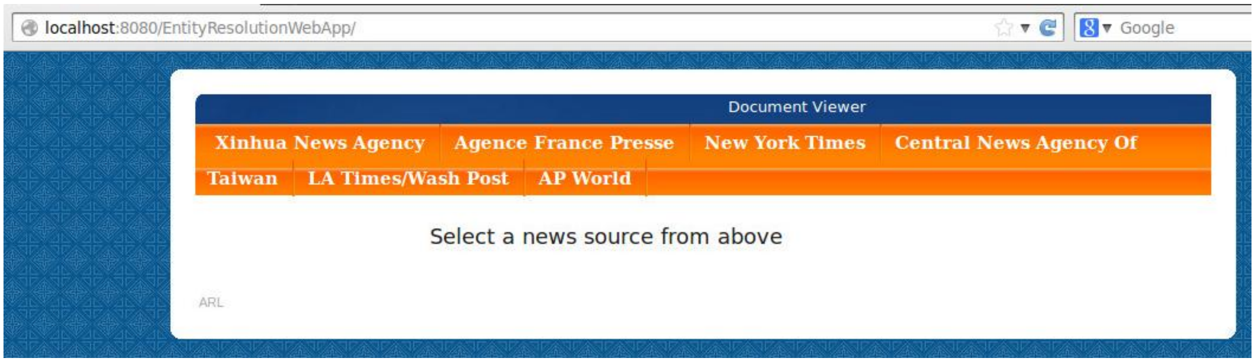
deploy EntityResolutionWebApp.war to <CATALINA_HOME>/webapps/EntityResolutionWebApp directory (figure 30). The console screen will show the deployment in action. The ERW Web Application is now installed and running.



```
Apr 3, 2013 10:09:53 AM org.apache.catalina.startup.Catalina start
INFO: Server startup in 867 ms
Apr 3, 2013 10:27:53 AM org.apache.catalina.startup.HostConfig deployWAR
INFO: Deploying web application archive /usr/local/apache-tomcat-7.0.37/webapps/EntityResolutionWebApp.war
```

Figure 30. Deploying EntityResolutionWebApp.war on Tomcat.

3. Browse to http://localhost:8080/EntityResolutionWebApp to use the ERW Web application (figure 31).



Select a news source from above

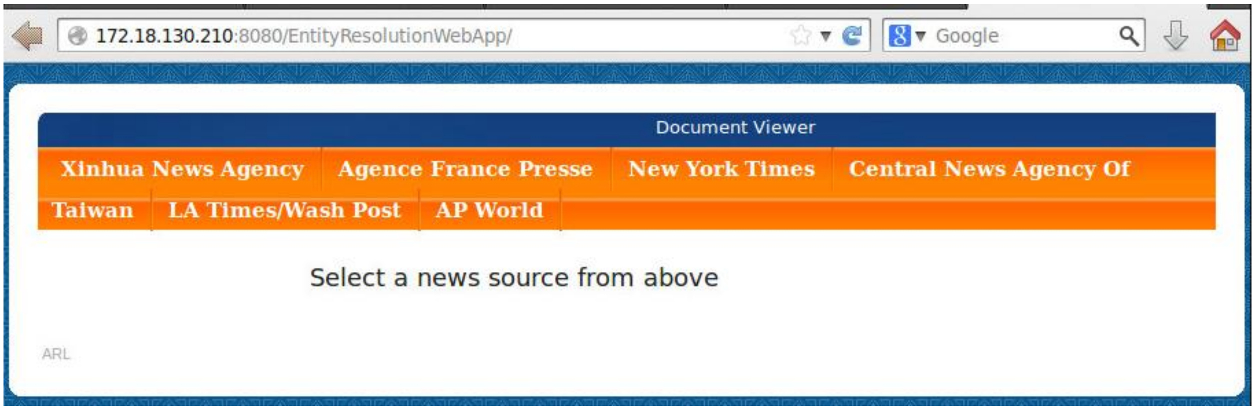
Figure 31. Starting page of the ERW Web Application.

The ERW Web application is now installed and ready to use. Please see the section 6 to learn how to use the ERW Web application.

6. ERW Web Application User Guide

The ERW Web application is a simple tool that demonstrates the features of ERW components and visualizes results of each step towards entity resolution. There five high-level steps for a single workflow: choose a document to process, perform entity extraction, perform candidate identification, perform entity resolution, and view resolved entity details. Each step is described in this section.

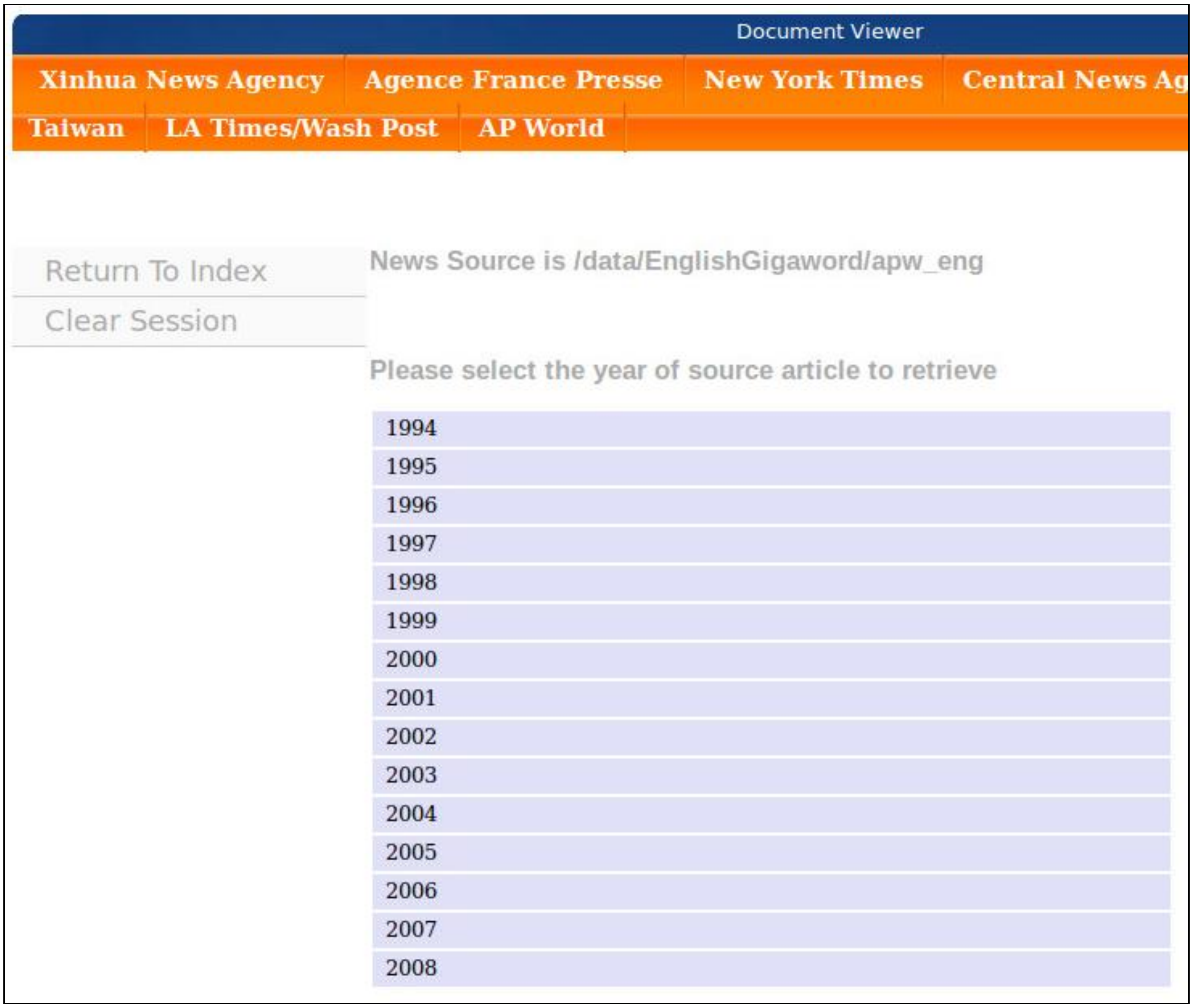
Load the ERW Web application by browsing to "[IP of ERDP-NN]:8080/EntityResolutionWebApp" (figure 32). At the time of writing this report, the IP address of ERDP-NN is 172.18.130.210. “EntityResolutionWebApp” is the name of the application use in Apache Tomcat.



Select a news source from above

Figure 32. Loading ERW Web application.

Once the ERW Web application is loaded, the user is asked to select one of the six news sources: Xinhua News Agency, Agency France Presse, New York Times, Central News Agency of Taiwan, LA Times/Wash Post, and AP World. These news sources are extracted from the English Gigaword corpus. (For more details on the English Gigaword, please see section 5 in this report). For this exercise, select “AP World” (figure 33).



News Source is /data/EnglishGigaword/apw_eng

Please select the year of source article to retrieve

1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008

Figure 33. List of available years in the AP World news source.

The list of years is displayed. These are the years the AP World articles were published. The English Gigaword contains AP World articles published between 1994 and 2008. For this exercise, select “2008”. The next two screens (figures 34 and 35) allow the user to specify month and a date in the selected year. For this exercise, select “December 31”.

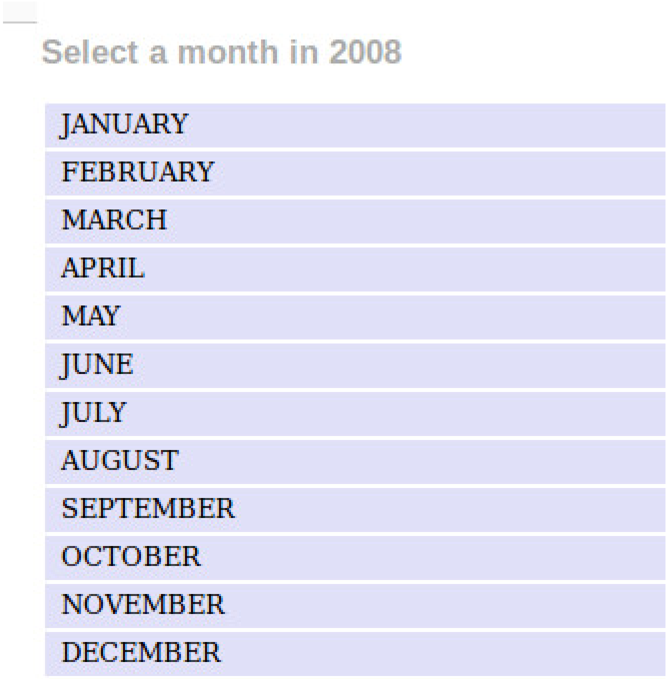


Figure 34. Month selection menu.

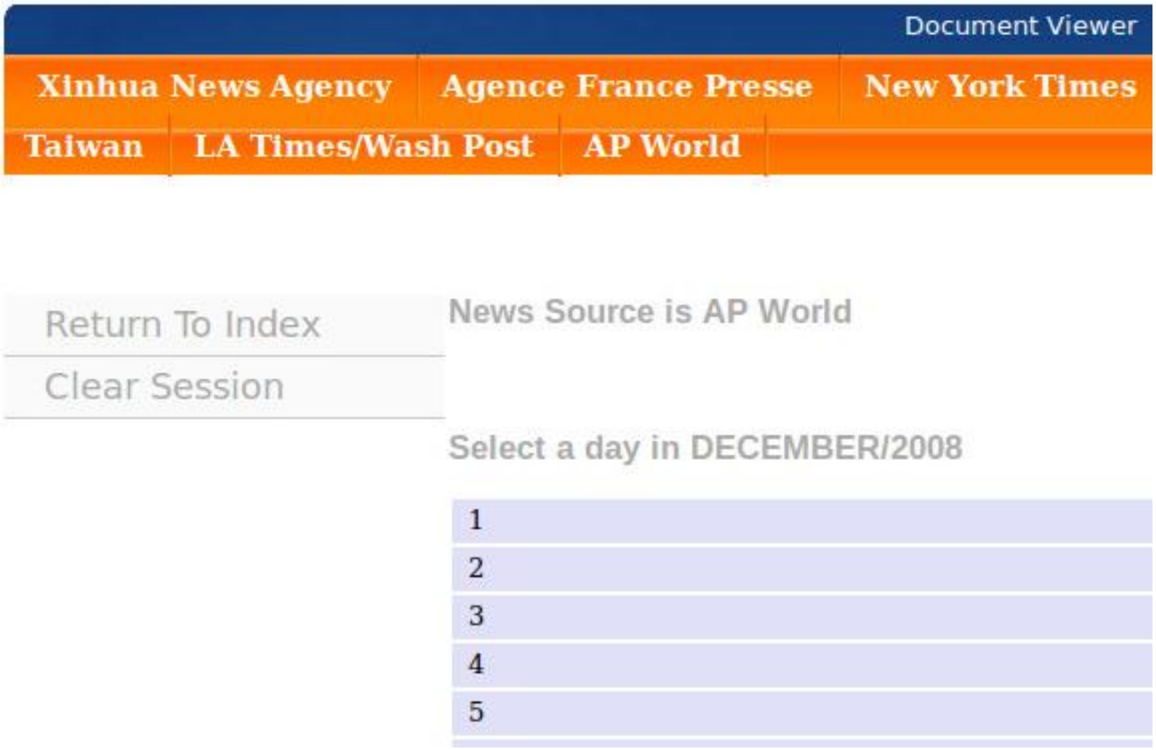
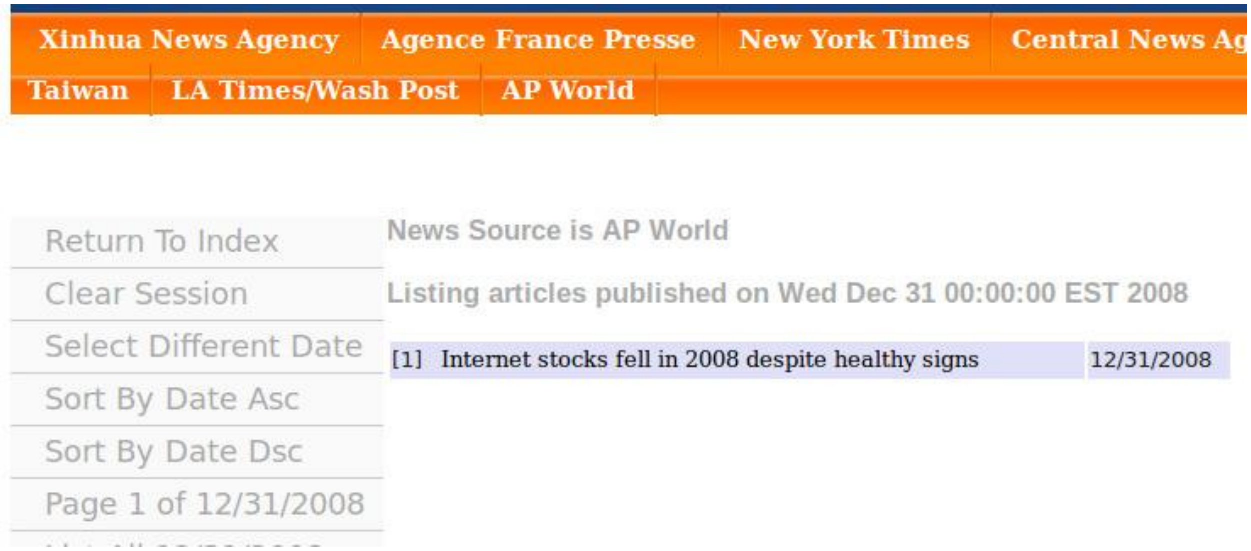


Figure 35. Date selection menu.

Once a news source and a date are selected, a list of articles is displayed. These are articles available in the English Gigaword that were published on the selected date by the selected news source. The number of articles in this list will vary from each news source and date selected. For example, there is only one AP World article available that was published on December 31, 2008 (figure 36), while there are 634 AP World articles available that were published on December 01, 2008 (figure 37). For this exercise, select the article “Internet stocks fell in 2008 despite healthy signs” (figure 38).

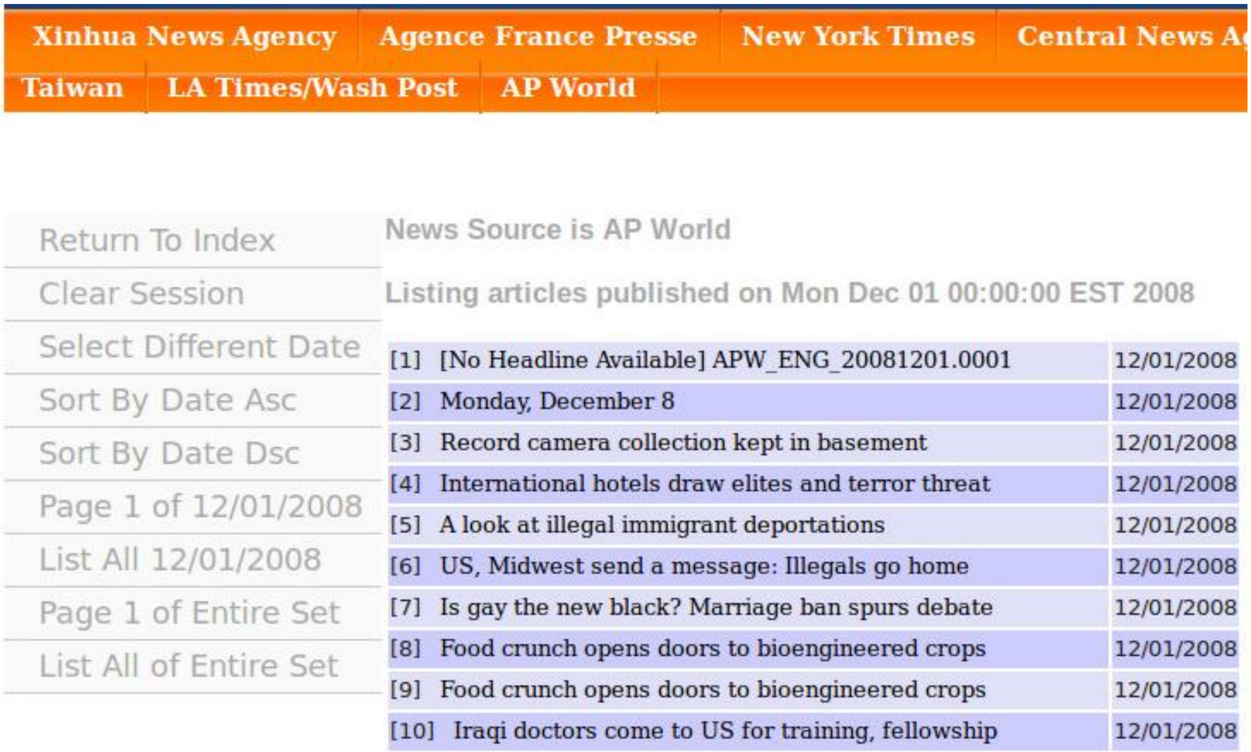


News Source is AP World

Listing articles published on Wed Dec 31 00:00:00 EST 2008

[1]	Internet stocks fell in 2008 despite healthy signs	12/31/2008

Figure 36. List of available articles published on December 31, 2008, by AP World.

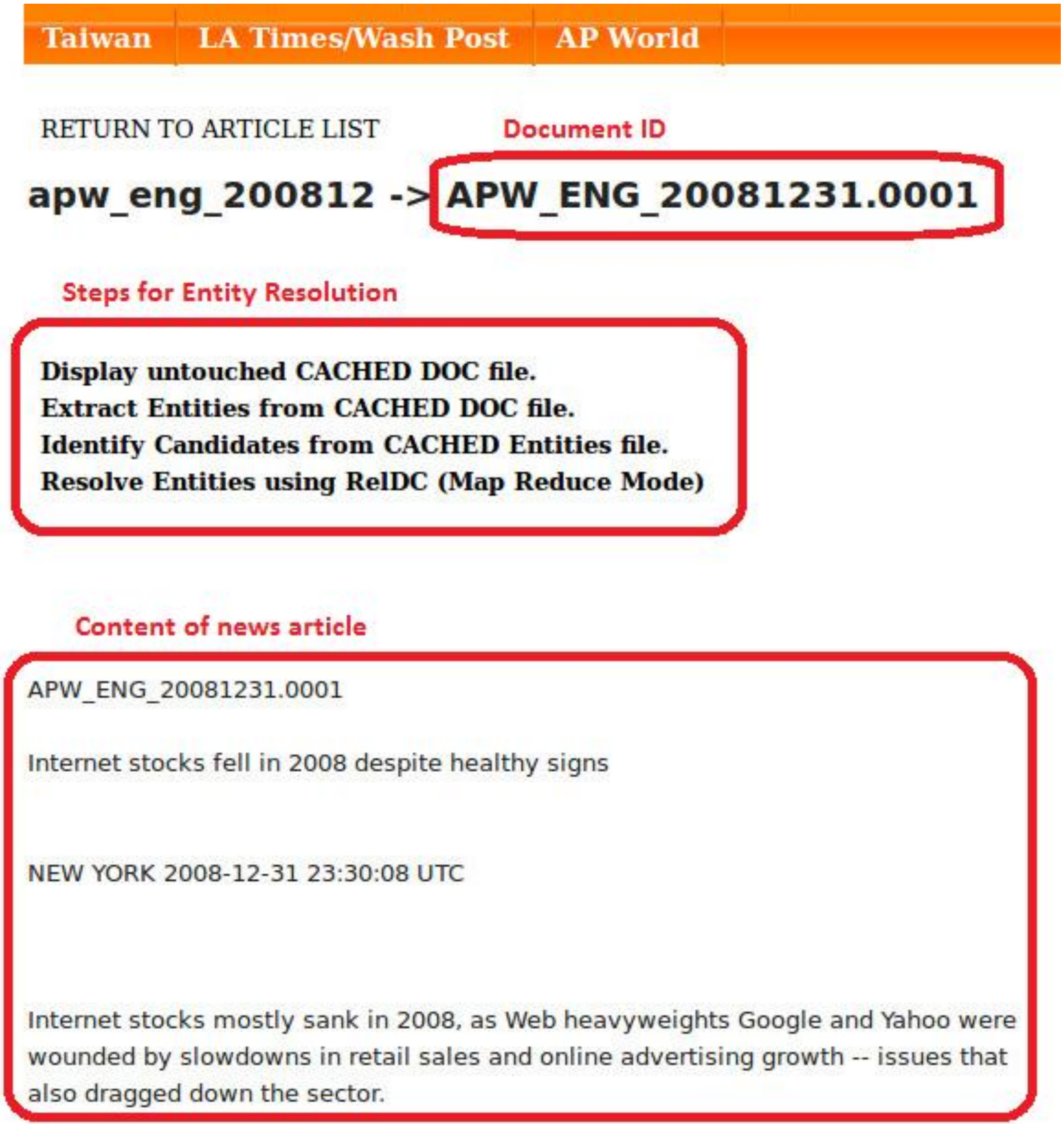


News Source is AP World

Listing articles published on Mon Dec 01 00:00:00 EST 2008

[1]	[No Headline Available] APW_ENG_20081201.0001	12/01/2008
[2]	Monday, December 8	12/01/2008
[3]	Record camera collection kept in basement	12/01/2008
[4]	International hotels draw elites and terror threat	12/01/2008
[5]	A look at illegal immigrant deportations	12/01/2008
[6]	US, Midwest send a message: Illegals go home	12/01/2008
[7]	Is gay the new black? Marriage ban spurs debate	12/01/2008
[8]	Food crunch opens doors to bioengineered crops	12/01/2008
[9]	Food crunch opens doors to bioengineered crops	12/01/2008
[10]	Iraqi doctors come to US for training, fellowship	12/01/2008

Figure 37. List of available articles published on December 01, 2008, by AP World.



Document ID

apw_eng_200812 -> APW_ENG_20081231.0001

Steps for Entity Resolution

Display untouched CACHED DOC file.
Extract Entities from CACHED DOC file.
Identify Candidates from CACHED Entities file.
Resolve Entities using RelDC (Map Reduce Mode)

Content of news article

APW_ENG_20081231.0001

Internet stocks fell in 2008 despite healthy signs

NEW YORK 2008-12-31 23:30:08 UTC

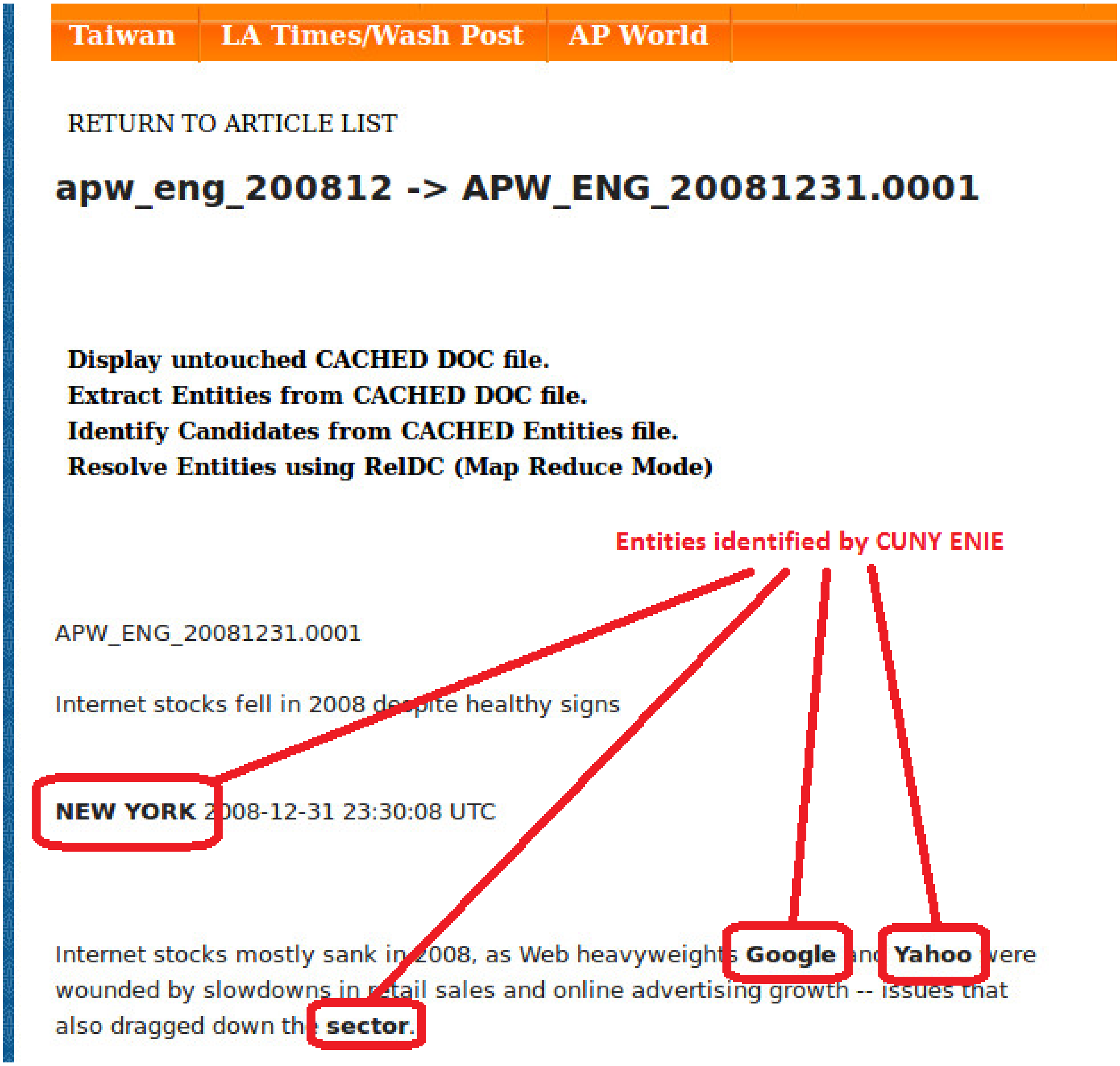
Internet stocks mostly sank in 2008, as Web heavyweights Google and Yahoo were wounded by slowdowns in retail sales and online advertising growth -- issues that also dragged down the sector.

Figure 38. Default view of the selected article.

The article view page displays the unique document ID at the top of the page, four steps for entity resolution under the document ID, and content of the article in the middle of the page. Article used in this exercise has a document ID “APW_ENG_20081231.0001” as shown in figure 38. The first three characters of the document ID represent the news source (e.g., “APW” represent “AP World”). The next three set of characters represent the language of the article (e.g., “ENG” represent “English”). Next group of characters represent the date of publication, formatted as YYYYMMDD (e.g., “20081231” represent “December 31, 2008”). Final four characters of the document ID is a counter of articles published on that day.

The current view of the article is displayed when “Display untouched CACHED DOC file” option is selected. This is the default and first step in the ERW process. As the title implies, this option simply displays the content of the selected article to the user. There is very little data processing required for this step, so this page will load immediately.

Click on the “Extract Entities from CACHED DOC file” link to proceed to the next step, Entity extraction page (figure 39). Loading the entity extraction page on this selected article for the first time will invoke CUNY ENIE and may take several minutes to process. After CUNY ENIE finishes processing the article, results will be cached so that subsequent requests for this page will load immediately. Identified entities are indicated in the body of the article with bold tags.



apw_eng_200812 -> APW_ENG_20081231.0001

Display untouched CACHED DOC file.
Extract Entities from CACHED DOC file.
Identify Candidates from CACHED Entities file.
Resolve Entities using RelDC (Map Reduce Mode)

Entities identified by CUNY ENIE

APW_ENG_20081231.0001

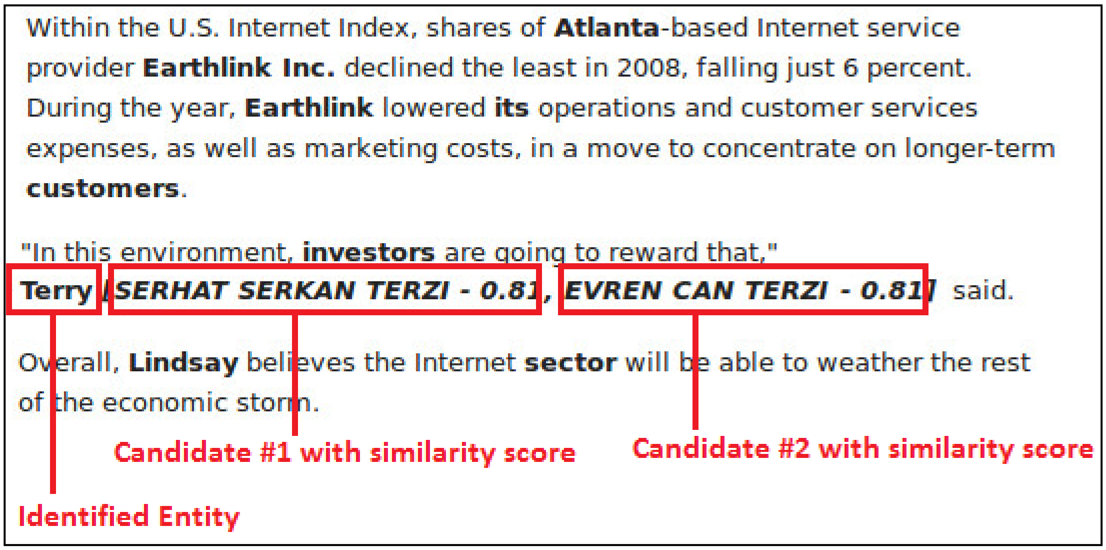
Internet stocks fell in 2008 despite healthy signs

NEW YORK 2008-12-31 23:30:08 UTC

Internet stocks mostly sank in 2008, as Web heavyweights **Google** and **Yahoo** were wounded by slowdowns in retail sales and online advertising growth -- issues that also dragged down the **sector**.

Figure 39. Entity extraction result page.

Click on the “Identify Candidates from CACHED Entities file” link to proceed to the next step, candidate identification page (figure 40). Again, initial load of candidate identification will take several minutes to process, and subsequent requests will load immediately. Candidates scoring higher than the threshold value are listed after the corresponding entity in the body of the article with bold tags. For example, the entity “Terry” has two possible candidates scoring above 0.80 — “Serhat Serkan Terzi” with a score 0.81 and “Evren Can Terzi” with a score 0.81. Note that these numeric values only measure string similarity to the entity “Terry”. These scores are not used in the entity resolution algorithm and a candidate with a lower string similarity score may be selected as the resolved entity.



Within the U.S. Internet Index, shares of **Atlanta**-based Internet service provider **Earthlink Inc.** declined the least in 2008, falling just 6 percent. During the year, **Earthlink** lowered **its** operations and customer services expenses, as well as marketing costs, in a move to concentrate on longer-term **customers**.

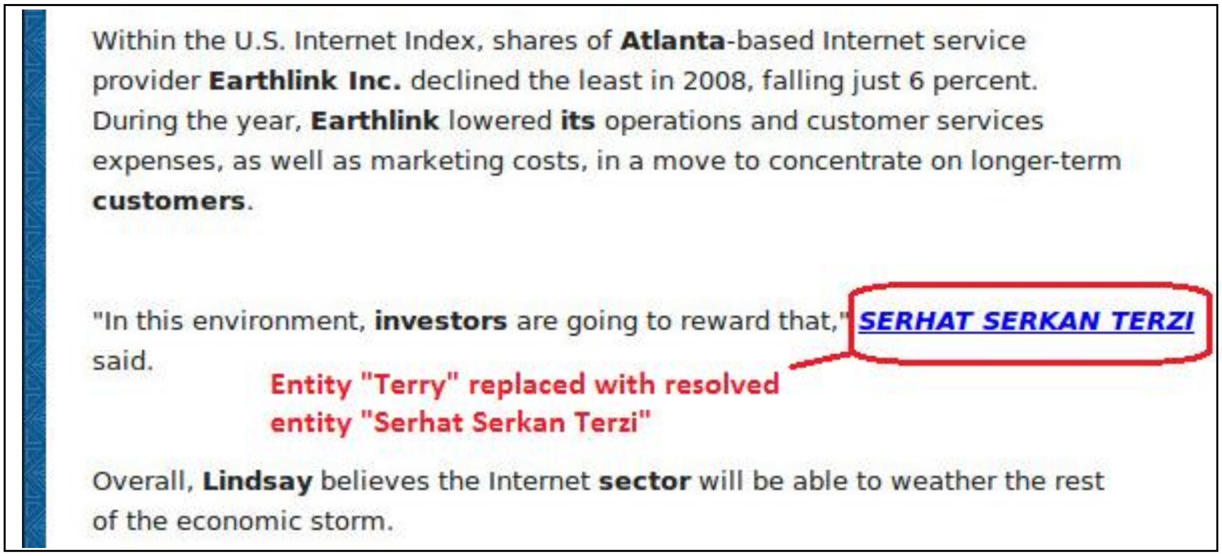
"In this environment, **investors** are going to reward that," **Terry** [***SERHAT SERKAN TERZI - 0.81***, ***EVREN CAN TERZI - 0.81***] said.

Overall, **Lindsay** believes the Internet **sector** will be able to weather the rest of the economic storm.

Candidate #1 with similarity score
Candidate #2 with similarity score
Identified Entity

Figure 40. Candidate identification result page.

Click on the “Resolve Entities using RelDC (Map Reduce Mode)” link to proceed to the next step, entity resolution page (figure 41). Loading the entity resolution page for the first time will invoke the Relationship Data Cleaning (RelDC) algorithm and may take several minutes to process. RelDC results will be cached so that subsequent requests for this page will load immediately. Resolved entities will replace the corresponding entities in the body of the article as a clickable link. In the previous example, the entity “Terry” had two candidates: “Serhat Serkan Terzi” and “Evren Can Terzi”. The RelDC algorithm calculated “Terry” and “Serhat Serkan Terzi”, in the entity base, are the same and entity resolution result page will display the article with the entity “Terry” replaced by the resolved entity “Serhat Serkan Terzi”.



Within the U.S. Internet Index, shares of **Atlanta**-based Internet service provider **Earthlink Inc.** declined the least in 2008, falling just 6 percent. During the year, **Earthlink** lowered **its** operations and customer services expenses, as well as marketing costs, in a move to concentrate on longer-term **customers**.

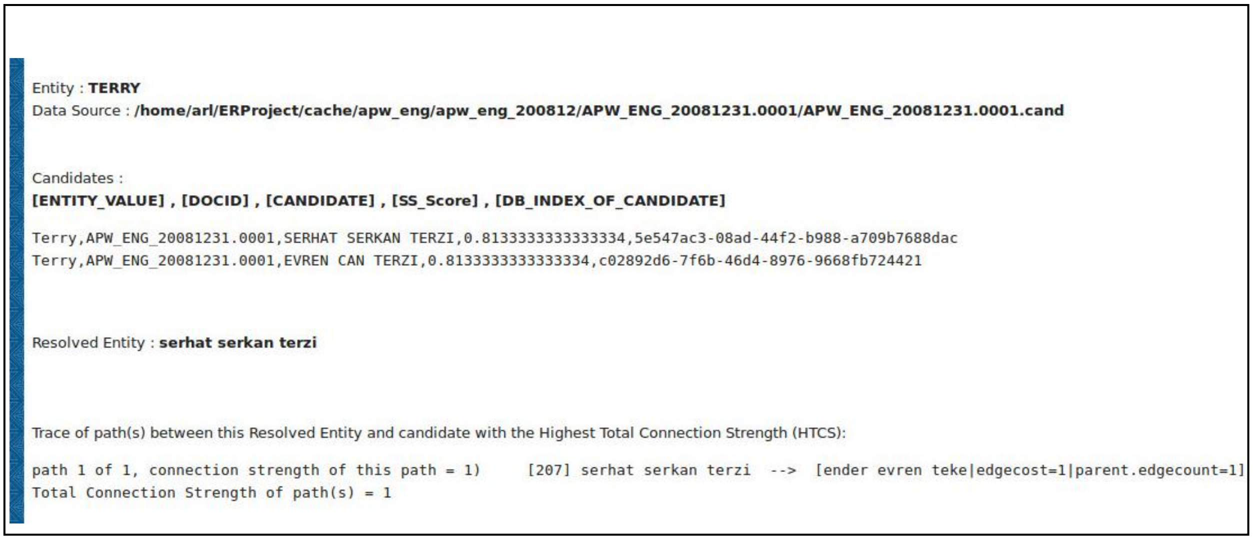
"In this environment, **investors** are going to reward that," ***SERHAT SERKAN TERZI*** said.

Entity "Terry" replaced with resolved entity "Serhat Serkan Terzi"

Overall, **Lindsay** believes the Internet **sector** will be able to weather the rest of the economic storm.

Figure 41. Resolved entities result page.

Clicking on a particular resolved entity will display details (e.g., source entity this resolved entity replaced, other potential candidates, path with the Highest Total Connect Strength, etc.) about that entry. For this exercise, click the resolved entity “SERHAT SERKAN TERZI”. The entity resolution details page (figure 42) shows the original identified entity (“Terry”) that was replaced by the resolved entity. The full path to the cached candidate identification result file is shown in the “Data Source” field, and candidate information that pertains to the identified entity is listed under the “Data Source” field. Each candidate detail includes entity value, document ID, candidate of interest, string similarity score of the candidate, and candidate UUID. Selected candidate is shown in the “Resolved Entity” label. Bottom of the page lists the path with the Highest Total Connection Strength (HTCS). In this example, it shows that there was a direct connection between the selected candidate (i.e., resolved entity) “Serhat Serkan Terzi” and another candidate (“Ender Evren Teke”) in this document. This path had the HTCS and therefore “Serhat Serkan Terzi” was selected as the resolved entity. Users can return to the resolved entities result page by clicking the “Return to Document” button at the top.



Entity : **TERRY**
Data Source : **/home/arl/ERProject/cache/apw_eng/apw_eng_200812/APW_ENG_20081231.0001/APW_ENG_20081231.0001.cand**

Candidates :
[ENTITY_VALUE] , [DOCID] , [CANDIDATE] , [SS_Score] , [DB_INDEX_OF_CANDIDATE]

```
Terry,APW_ENG_20081231.0001,SERHAT SERKAN TERZI,0.8133333333333334,5e547ac3-08ad-44f2-b988-a709b7688dac
Terry,APW_ENG_20081231.0001,EVREN CAN TERZI,0.8133333333333334,c02892d6-7f6b-46d4-8976-9668fb724421
```

Resolved Entity : **serhat serkan terzi**

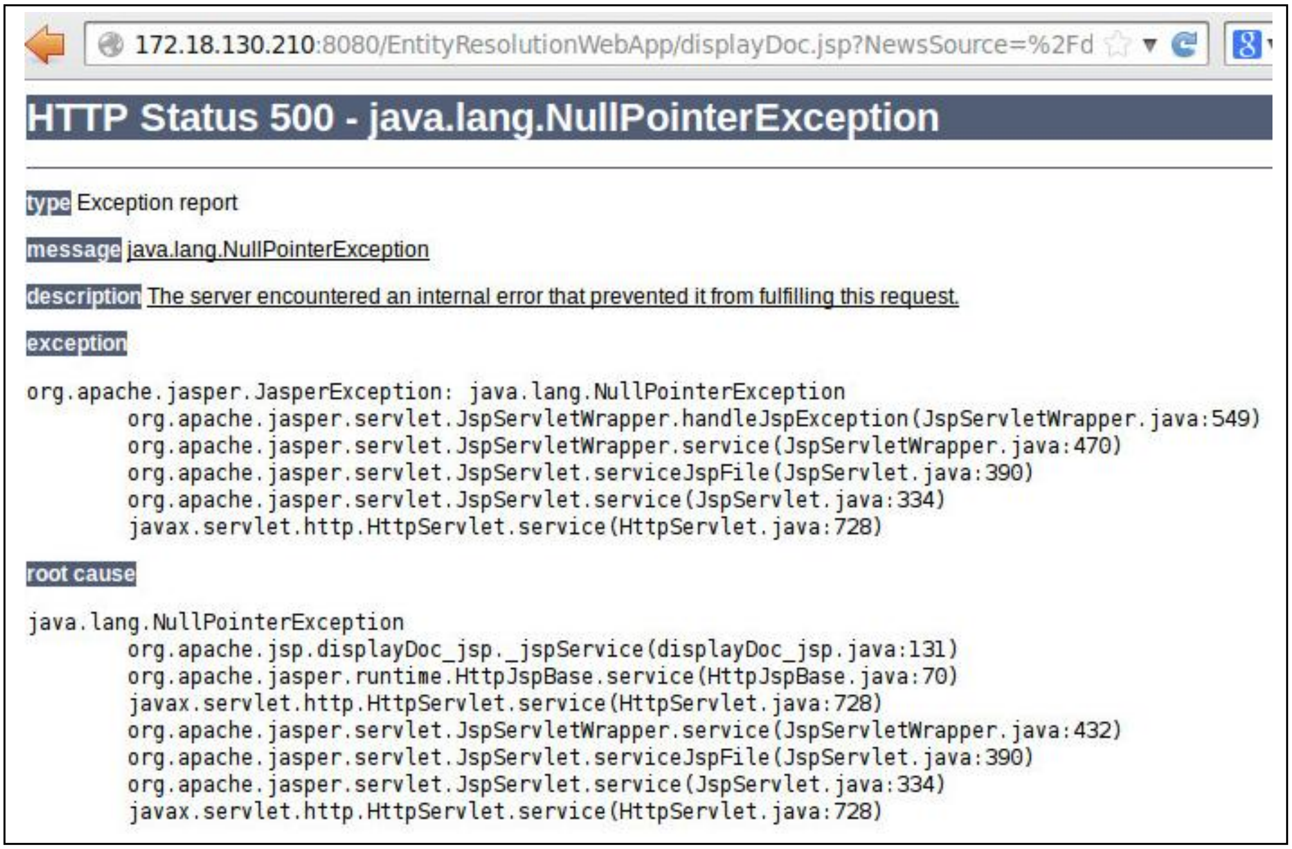
Trace of path(s) between this Resolved Entity and candidate with the Highest Total Connection Strength (HTCS):

```
path 1 of 1, connection strength of this path = 1)     [207] serhat serkan terzi  -->  [ender evren teke|edgecost=1|parent.edgecount=1]
Total Connection Strength of path(s) = 1
```

Figure 42. Resolved entity details page.

6.1 Caveat: “NullPointerException” Error in ERW Web Application

After 30 min of activity, the ERW Web application’s session will expire automatically. The user will see a “NullPointerException” error as illustrated in figure 43.



172.18.130.210:8080/EntityResolutionWebApp/displayDoc.jsp?NewsSource=%2Fd

HTTP Status 500 - java.lang.NullPointerException

type Exception report

message java.lang.NullPointerException

description The server encountered an internal error that prevented it from fulfilling this request.

exception

```
org.apache.jasper.JasperException: java.lang.NullPointerException
	org.apache.jasper.servlet.JspServletWrapper.handleJspException(JspServletWrapper.java:549)
	org.apache.jasper.servlet.JspServletWrapper.service(JspServletWrapper.java:470)
	org.apache.jasper.servlet.JspServlet.serviceJspFile(JspServlet.java:390)
	org.apache.jasper.servlet.JspServlet.service(JspServlet.java:334)
	javax.servlet.http.HttpServlet.service(HttpServlet.java:728)
```

root cause

```
java.lang.NullPointerException
	org.apache.jsp.displayDoc_jsp._jspService(displayDoc_jsp.java:131)
	org.apache.jasper.runtime.HttpJspBase.service(HttpJspBase.java:70)
	javax.servlet.http.HttpServlet.service(HttpServlet.java:728)
	org.apache.jasper.servlet.JspServletWrapper.service(JspServletWrapper.java:432)
	org.apache.jasper.servlet.JspServlet.serviceJspFile(JspServlet.java:390)
	org.apache.jasper.servlet.JspServlet.service(JspServlet.java:334)
	javax.servlet.http.HttpServlet.service(HttpServlet.java:728)
```

Figure 43. “NullPointerException” error on ERW Web application.

When this error is encountered, click the browser's back button until left column menus appear. Click on the "Clear Session" link to resolve this error and renew the session (figure 44).

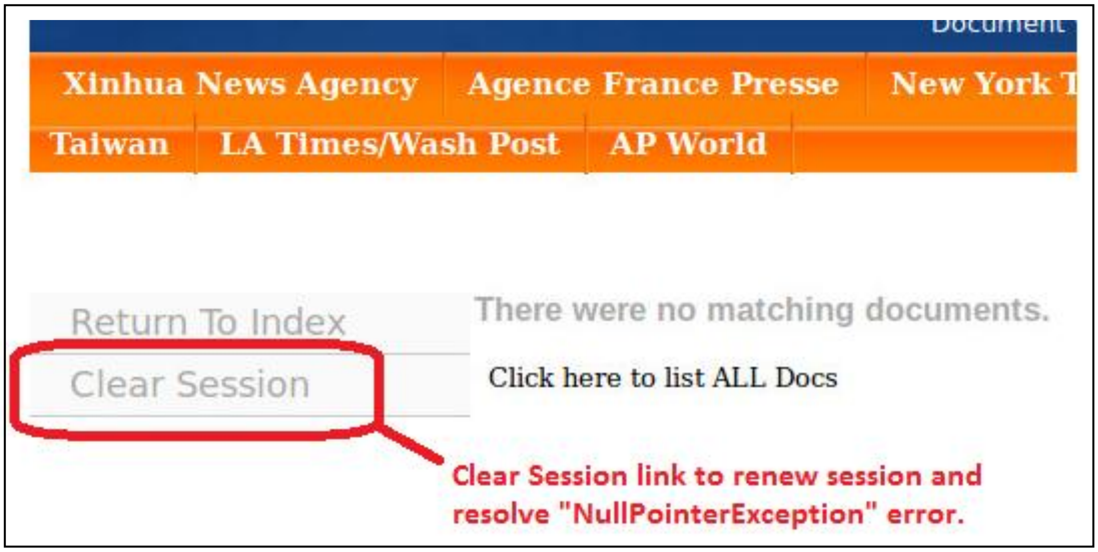

Document
Xinhua News Agency
Agence France Presse
New York T
Taiwan
LA Times/Wash Post
AP World
Return To Index
Clear Session
There were no matching documents.
Click here to list ALL Docs
Clear Session link to renew session and resolve "NullPointerException" error.

Figure 44. "Clear Session" link for resolving the "NullPointerException" error.

7. Conclusion

ERW is an easy-to-use system to test the RelDC entity resolution algorithm and other components required for RelDC, while shielding the users from complexities of working with many components needed by RelDC. In this report, we described the installation, configuration, and use of this system. Once ERW has been successfully setup, users can choose to use the built-in components shipped with ERW or swap components to compare the results.

Appendix A. Installing Global Graph on ERDP-WIN7 (Windows 7 OS)

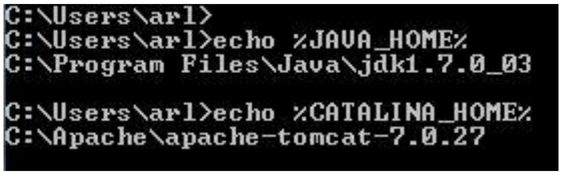
Installation document included with Global Graph distribution is targeted for Windows OS. It is possible to install Global Graph and its components (PostgreSQL, PostGIS, Java, and Apache Tomcat) on a Linux OS, but it requires lot of effort troubleshooting configuration issues. ERW will install Global Graph on a Windows OS to simplify the installation process. This appendix documents the steps taken to install Global Graph 1.4.6 on ERDP-WIN7 (Windows 7 OS). These steps are for reference only and configured with very loose security measures for demonstration purpose. Please contact Potomac Fusion Inc. (PFI) for any installation problems.

Note: Some Global Graph properties files specify IP addresses.

1. Install Java JDK and Apache Tomcat:

   - Install Java JDK with default settings and verify “JAVA_HOME” environment variable is set, and add “%JAVA_HOME%\bin” to the system path.

   - Install Apache Tomcat and verify “CATALINA_HOME” environment variable (figure A-1) is set.



```
C:\Users\arl>
C:\Users\arl>echo %JAVA_HOME%
C:\Program Files\Java\jdk1.7.0_03

C:\Users\arl>echo %CATALINA_HOME%
C:\Apache\apache-tomcat-7.0.27
```

Figure A-1. Java and Tomcat environment variables.

2. Install PostgreSQL:

   - Download postgresql-9.1.3-1-windows.exe from http://get.enterprisedb.com/postgresql/postgresql-9.1.3-1-windows.exe.

     These setting were used for the installation wizard:

     Installation Directory: C:\Program Files\PostgreSQL\9.1

     Data Directory: C:\Program Files\PostgreSQL\9.1\data

     superuser (postgres), service account (postgres) password : "password"

     Port #: 5432

     Add an environment variable “PG_HOME” and set the value as “C:\Program Files\PostgreSQL\9.1” (do not include double quotes).

3. Install PostGIS:

   - Download postgis-pg91-setup-1.5.3-2.exe from http://www.postgis.org/download/windows/pg91/postgis-pg91-setup-1.5.3-2.exe.

     Follow the PostGIS installation wizard to complete the installation.

     Select all components for installation.

     Enter “postgis” when prompted for the database name

     Select “Yes” when prompted to install “shp2pgsql” plugin.

4. Configure and setup PostgreSQL for GlobalGraph:

   An installation script is included with the Global Graph distribution. Copy *<DISTRIBUTION_HOME>/GlobalGraph/GG-1.4.6* directory to ERDP-Win7 on *C:\Program Files\Global Graph\1.4.6*, which is referred to as *<GG_HOME>* in this report.

   - Run the script found in <GG_HOME>\ globalgraph-dist-1.4.6-final\schema-ddl\postgresSetup.bat. This script will set up Postgres admin user, Global Graph database user, passwords, create databases, and populate the databases. Enter the following when prompted:

DB Admin Username:	postgres
DB Admin PWD:	password
GlobalGraph App User:	gguser
GlobalGraph App PWD:	password

   - Restart the Postgres service using the Windows Services Manager.

5. Enable remote connection on PostgresSQL:

   Remote connection must be enabled so that Hadoop Cluster data nodes can connect to the PostgreSQL server to query the Global Graph. Remote connection is disabled by default. Follow these steps to enable remote connection:

   - Verify <POSTGRESQL_HOME>/data/postgresql.conf contains the line “listen_addresses = '*'”

- Edit <POSTGRESQL_HOME>/data/pg_hba.conf co and add IP of servers that will be allowed to connect to PostgreSQL. For this demonstration PostgreSQL will be set to allow all connections. Add the following line to pg_hba.conf file:

  ```
  host all all  0.0.0.0/0 md5
  ```

  This is an access control rule that lets anyone login from any address if a valid password is provided (the md5 keyword).

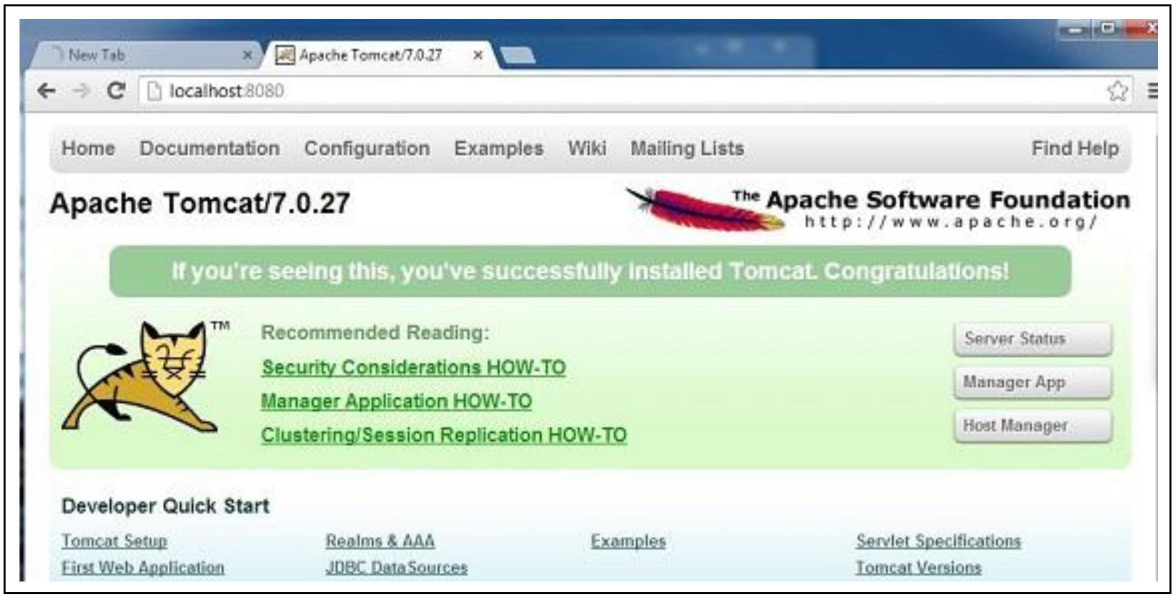
- Restart the Postgres service using the Windows Services Manager.
- PostgreSQL will now allow remote connections. Verify by connecting from a remote machine.

6. Install Global Graph REST Services:

- Verify Apache Tomcat is configured correctly by starting Tomcat:

  ```
  %CATALINA_HOME%\bin\startup.bat
  ```

- Open browser to http://localhost:8080. If a welcome page (figure A-2) appears, then Tomcat is correctly installed.



localhost:8080

Apache Tomcat/7.0.27

If you're seeing this, you've successfully installed Tomcat. Congratulations!

Figure A-2. Apache Tomcat welcome page.

- Extract *<GG_HOME>/globalgraph-dist-1.4.6-final/dist/rest-service-1.4.6-A2SF-RC2.war* (which contains "META-INF" and "WEB-INF" directories) to *<CATALINA_HOME>/webapps/gg directory*.

- Copy postgres*.jar and postgis*.jar (two files) from *<DISTRIBUTION_HOME>/GlobalGraph/GG-1.4.6/globalgraph-dist-1.4.6-final/dist/all-lib to <CATALINA_HOME>/lib* directory.
- Modify *<CATALINA_HOME>/webapps/gg/WEB-INF/classes/jdbc.properties* file. Update these 3 properties with values set during PostgreSQL installation:

  jdbc.url=jdbc:postgresql://localhost:5432/globalgraph?useUnicode=true&amp;characterEncoding=utf-8

  jdbc.username=gguser

  jdbc.password=password
- Modify *<CATALINA_HOME>/webapps/gg/META-INF/context.xml* file. Update these three properties with values set during PostgreSQL installation:

  username="gguser"

  password="password"

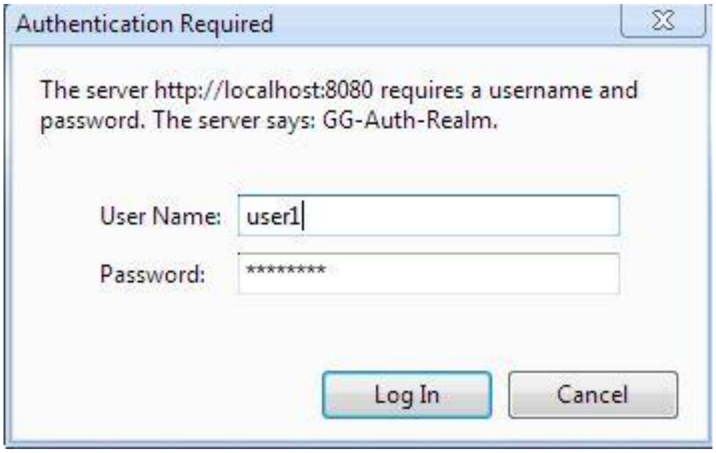
  url="jdbc:postgresql://localhost:5432/globalgraph?useUnicode=true&amp;characterEncoding=utf-8"
- Restart Tomcat.

7. Test Global Graph REST Services:

   - Test the Global Graph REST Services by requesting a person JSON using a browser. Open a browser to http://localhost:8080/gg/entity/Person/00914273-e995-42f0-a428-77170495311a. "00914273-e995-42f0-a428-77170495311a" is a UID of a random person found in the globalgraph database.
   - When prompted for user name and password (figure A-3), use the Global Graph's preloaded user name and password:

     User Name: user1

     Password: ggsecret



Authentication Required

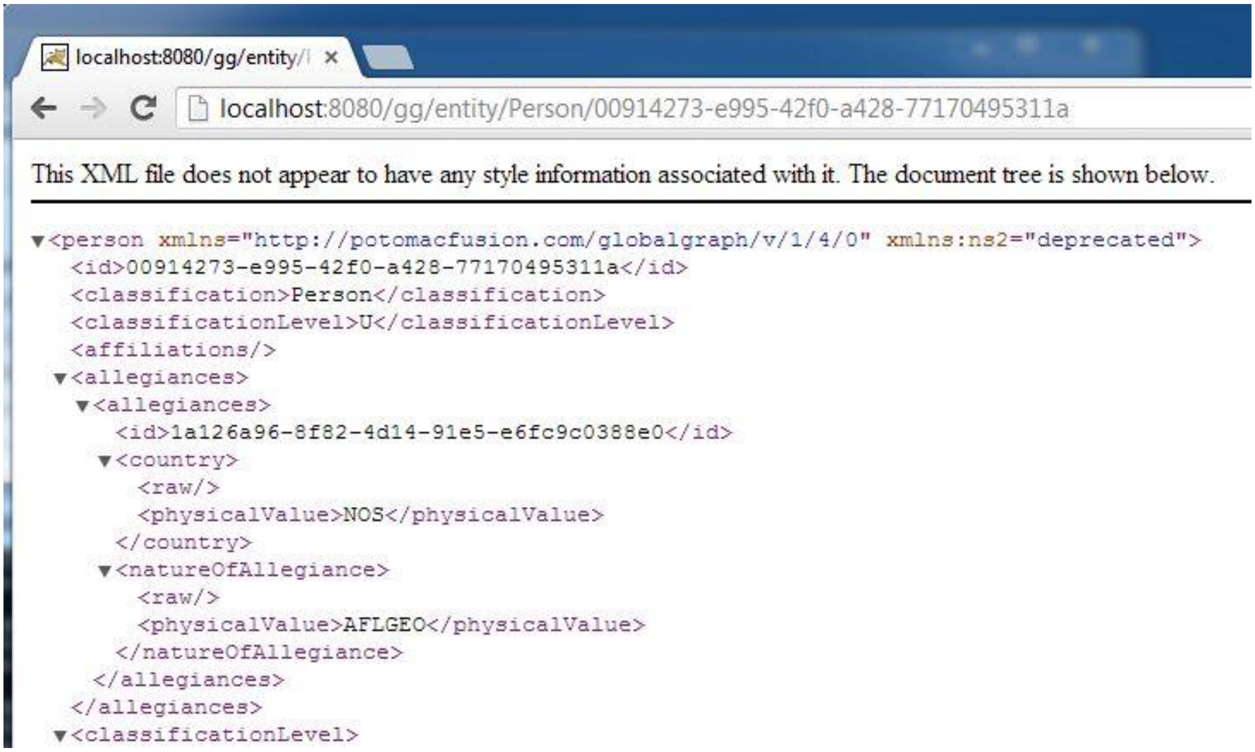
The server http://localhost:8080 requires a username and password. The server says: GG-Auth-Realm.

User Name: user1

Password: ********

Figure A-3. Global Graph REST Services authentication page.

- Once the user's credentials are accepted, JSON result (figure A-4) is displayed. Global Graph is now installed on ERDP-Win7, and the Global Graph REST Services are now accessible from the Hadoop Cluster data nodes.



localhost:8080/gg/entity/Person/00914273-e995-42f0-a428-77170495311a

This XML file does not appear to have any style information associated with it. The document tree is shown below.

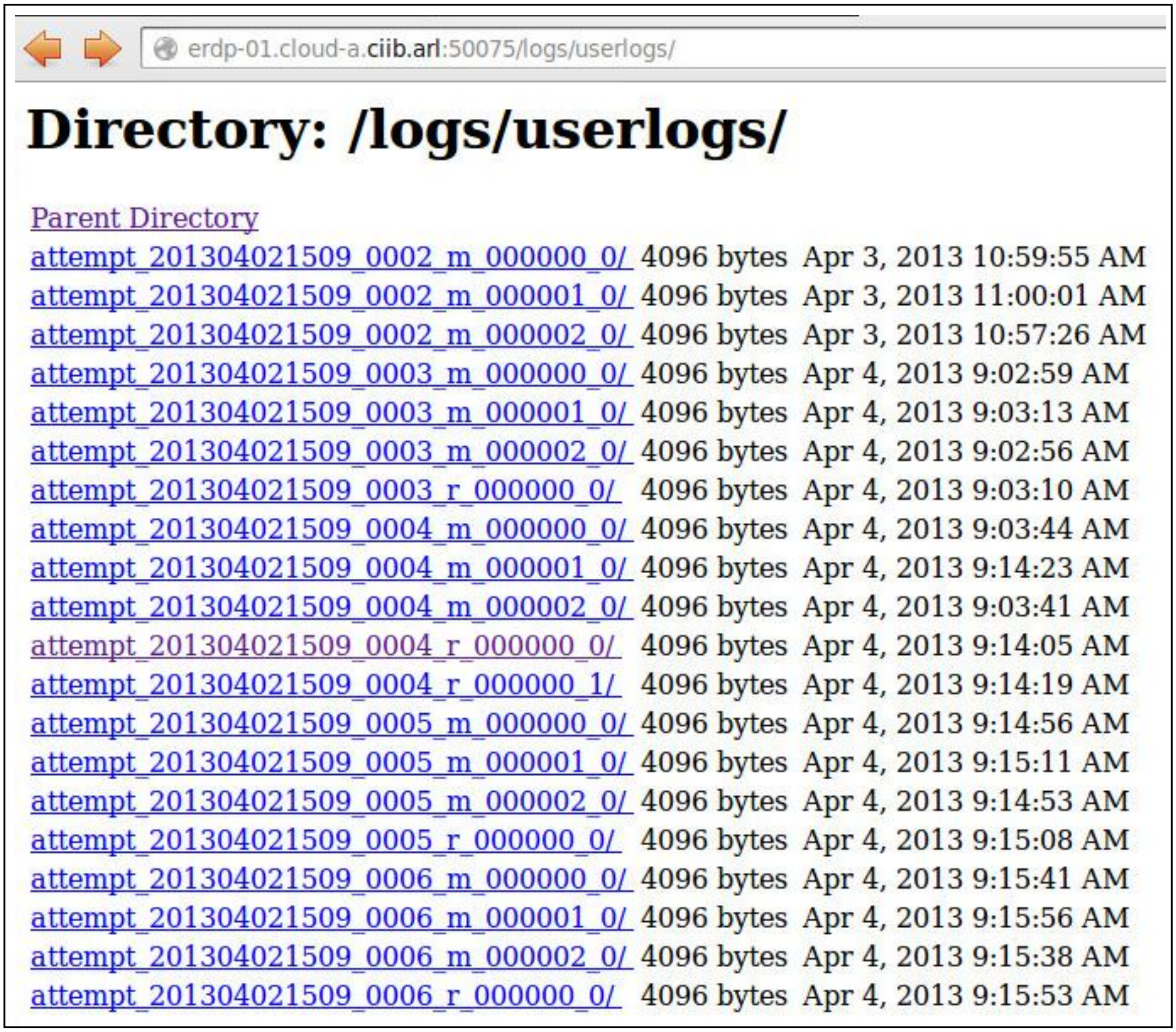
```
▼<person xmlns="http://potomacfusion.com/globalgraph/v/1/4/0" xmlns:ns2="deprecated">
   <id>00914273-e995-42f0-a428-77170495311a</id>
   <classification>Person</classification>
   <classificationLevel>U</classificationLevel>
   <affiliations/>
 ▼<allegiances>
   ▼<allegiances>
      <id>1a126a96-8f82-4d14-91e5-e6fc9c0388e0</id>
     ▼<country>
        <raw/>
        <physicalValue>NOS</physicalValue>
      </country>
     ▼<natureOfAllegiance>
        <raw/>
        <physicalValue>AFLGEO</physicalValue>
      </natureOfAllegiance>
    </allegiances>
   </allegiances>
 ▼<classificationLevel>
```

Figure A-4. JSON string for a person entity.

INTENTIONALLY LEFT BLANK.

Appendix B. How to Search for Debug Messages in Data Nodes

When a map-reduce PIG script runs, the job is transferred and processed on a separate data node. The user does not know which data node will process that job. In addition, the data node that processes the job creates many log directories (with names that do not provide much value in searching for the correct log file) (figure B-1) and files ("stderr", "stdout", and "syslog") (figure B-2) for that single job. Even if the user knew which data node was processing the job, it will be time consuming searching each directory to identify the log file of interest.



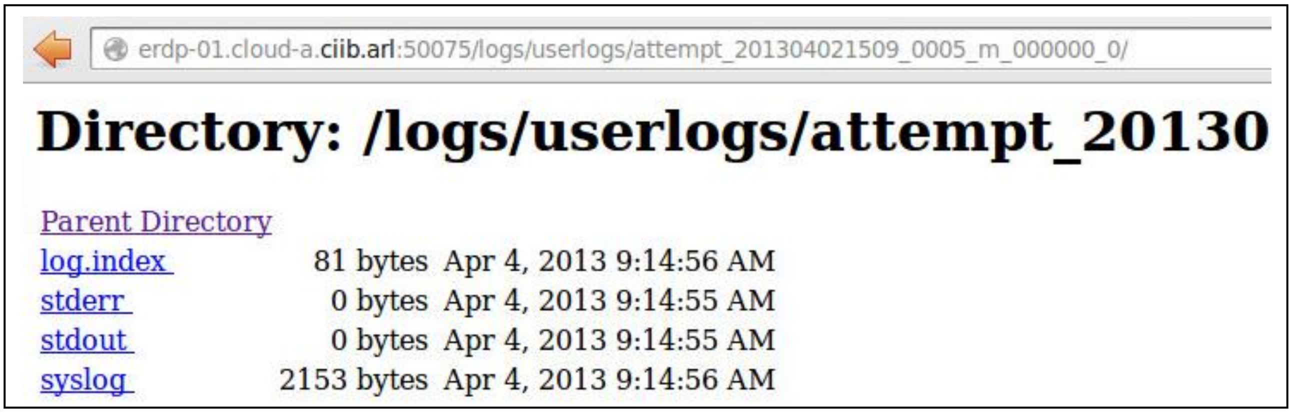
erdp-01.cloud-a.ciib.arl:50075/logs/userlogs/

Directory: /logs/userlogs/

Parent Directory

attempt_201304021509_0002_m_000000_0/	4096 bytes	Apr 3, 2013 10:59:55 AM
attempt_201304021509_0002_m_000001_0/	4096 bytes	Apr 3, 2013 11:00:01 AM
attempt_201304021509_0002_m_000002_0/	4096 bytes	Apr 3, 2013 10:57:26 AM
attempt_201304021509_0003_m_000000_0/	4096 bytes	Apr 4, 2013 9:02:59 AM
attempt_201304021509_0003_m_000001_0/	4096 bytes	Apr 4, 2013 9:03:13 AM
attempt_201304021509_0003_m_000002_0/	4096 bytes	Apr 4, 2013 9:02:56 AM
attempt_201304021509_0003_r_000000_0/	4096 bytes	Apr 4, 2013 9:03:10 AM
attempt_201304021509_0004_m_000000_0/	4096 bytes	Apr 4, 2013 9:03:44 AM
attempt_201304021509_0004_m_000001_0/	4096 bytes	Apr 4, 2013 9:14:23 AM
attempt_201304021509_0004_m_000002_0/	4096 bytes	Apr 4, 2013 9:03:41 AM
attempt_201304021509_0004_r_000000_0/	4096 bytes	Apr 4, 2013 9:14:05 AM
attempt_201304021509_0004_r_000000_1/	4096 bytes	Apr 4, 2013 9:14:19 AM
attempt_201304021509_0005_m_000000_0/	4096 bytes	Apr 4, 2013 9:14:56 AM
attempt_201304021509_0005_m_000001_0/	4096 bytes	Apr 4, 2013 9:15:11 AM
attempt_201304021509_0005_m_000002_0/	4096 bytes	Apr 4, 2013 9:14:53 AM
attempt_201304021509_0005_r_000000_0/	4096 bytes	Apr 4, 2013 9:15:08 AM
attempt_201304021509_0006_m_000000_0/	4096 bytes	Apr 4, 2013 9:15:41 AM
attempt_201304021509_0006_m_000001_0/	4096 bytes	Apr 4, 2013 9:15:56 AM
attempt_201304021509_0006_m_000002_0/	4096 bytes	Apr 4, 2013 9:15:38 AM
attempt_201304021509_0006_r_000000_0/	4096 bytes	Apr 4, 2013 9:15:53 AM

Figure B-1. Log directories on a data node.



erdp-01.cloud-a.ciib.arl:50075/logs/userlogs/attempt_201304021509_0005_m_000000_0/

Directory: /logs/userlogs/attempt_20130

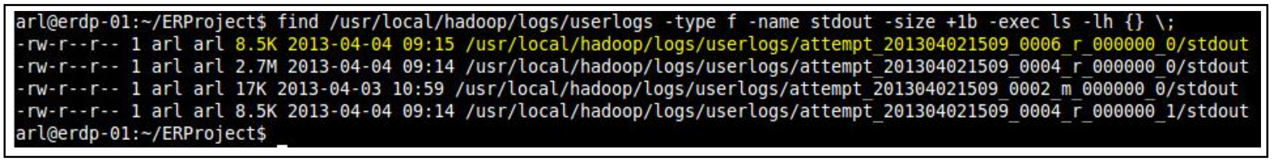
```
Parent Directory
log.index     81 bytes  Apr 4, 2013 9:14:56 AM
stderr         0 bytes  Apr 4, 2013 9:14:55 AM
stdout         0 bytes  Apr 4, 2013 9:14:55 AM
syslog      2153 bytes  Apr 4, 2013 9:14:56 AM
```

Figure B-2. Files in the log directory.

ERW components log their debug messages to the “stdout” file. The user needs to find a “stdout” file (figure B-3) that is not empty and was created after the PIG script ran. The following command searches the Hadoop logs directory and lists “stdout” file that is not empty:

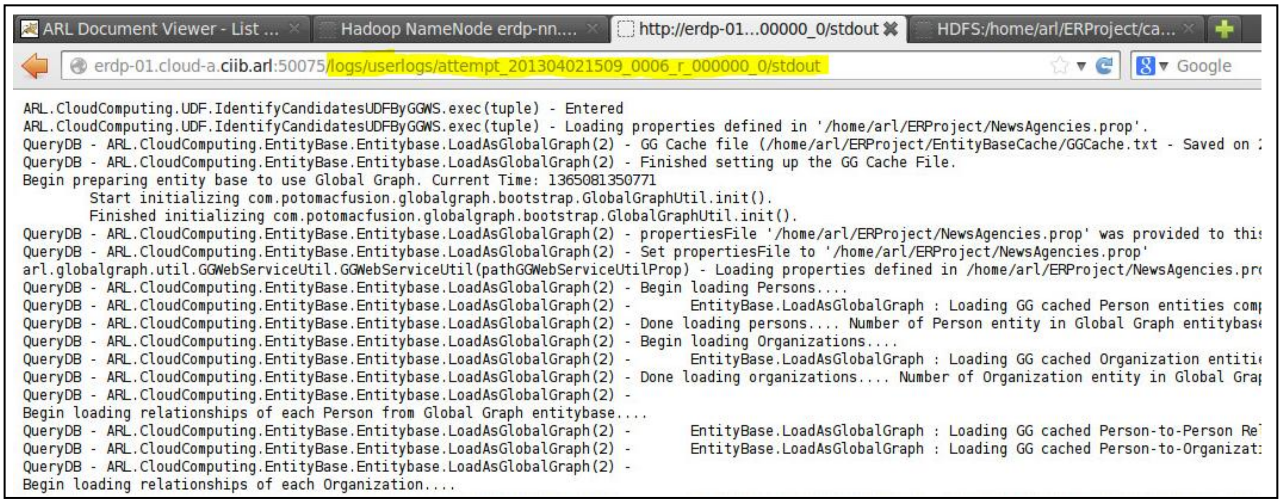
find /usr/local/hadoop/logs/userlogs -type f -name stdout -size +1b -exec ls -lh {} \;

The log file of interest (figure B-4) is most likely the file with the most recent time stamp. The debug statements in the log file will be very helpful with identifying the problem.



```
arl@erdp-01:~/ERProject$ find /usr/local/hadoop/logs/userlogs -type f -name stdout -size +1b -exec ls -lh {} \;
-rw-r--r-- 1 arl arl 8.5K 2013-04-04 09:15 /usr/local/hadoop/logs/userlogs/attempt_201304021509_0006_r_000000_0/stdout
-rw-r--r-- 1 arl arl 2.7M 2013-04-04 09:14 /usr/local/hadoop/logs/userlogs/attempt_201304021509_0004_r_000000_0/stdout
-rw-r--r-- 1 arl arl 17K 2013-04-03 10:59 /usr/local/hadoop/logs/userlogs/attempt_201304021509_0002_m_000000_0/stdout
-rw-r--r-- 1 arl arl 8.5K 2013-04-04 09:14 /usr/local/hadoop/logs/userlogs/attempt_201304021509_0004_r_000000_1/stdout
arl@erdp-01:~/ERProject$ _
```

Figure B-3. Identifying “stdout” log file of interest.



ARL Document Viewer - List ...
Hadoop NameNode erdp-nn....
http://erdp-01...00000_0/stdout
HDFS:/home/arl/ERProject/ca...
Google

erdp-01.cloud-a.ciib.arl:50075/logs/userlogs/attempt_201304021509_0006_r_000000_0/stdout

```
ARL.CloudComputing.UDF.IdentifyCandidatesUDFByGGWS.exec(tuple) - Entered
ARL.CloudComputing.UDF.IdentifyCandidatesUDFByGGWS.exec(tuple) - Loading properties defined in '/home/arl/ERProject/NewsAgencies.prop'.
QueryDB - ARL.CloudComputing.EntityBase.Entitybase.LoadAsGlobalGraph(2) - GG Cache file (/home/arl/ERProject/EntityBaseCache/GGCache.txt - Saved on 
QueryDB - ARL.CloudComputing.EntityBase.Entitybase.LoadAsGlobalGraph(2) - Finished setting up the GG Cache File.
Begin preparing entity base to use Global Graph. Current Time: 1365081350771
        Start initializing com.potomacfusion.globalgraph.bootstrap.GlobalGraphUtil.init().
        Finished initializing com.potomacfusion.globalgraph.bootstrap.GlobalGraphUtil.init().
QueryDB - ARL.CloudComputing.EntityBase.Entitybase.LoadAsGlobalGraph(2) - propertiesFile '/home/arl/ERProject/NewsAgencies.prop' was provided to this
QueryDB - ARL.CloudComputing.EntityBase.Entitybase.LoadAsGlobalGraph(2) - Set propertiesFile to '/home/arl/ERProject/NewsAgencies.prop'
arl.globalgraph.util.GGWebServiceUtil.GGWebServiceUtil(pathGGWebServiceUtilProp) - Loading properties defined in /home/arl/ERProject/NewsAgencies.pro
QueryDB - ARL.CloudComputing.EntityBase.Entitybase.LoadAsGlobalGraph(2) - Begin loading Persons....
QueryDB - ARL.CloudComputing.EntityBase.Entitybase.LoadAsGlobalGraph(2) -       EntityBase.LoadAsGlobalGraph : Loading GG cached Person entities comp
QueryDB - ARL.CloudComputing.EntityBase.Entitybase.LoadAsGlobalGraph(2) - Done loading persons.... Number of Person entity in Global Graph entitybase
QueryDB - ARL.CloudComputing.EntityBase.Entitybase.LoadAsGlobalGraph(2) - Begin loading Organizations....
QueryDB - ARL.CloudComputing.EntityBase.Entitybase.LoadAsGlobalGraph(2) -       EntityBase.LoadAsGlobalGraph : Loading GG cached Organization entitie
QueryDB - ARL.CloudComputing.EntityBase.Entitybase.LoadAsGlobalGraph(2) - Done loading organizations.... Number of Organization entity in Global Grap
QueryDB - ARL.CloudComputing.EntityBase.Entitybase.LoadAsGlobalGraph(2) -
Begin loading relationships of each Person from Global Graph entitybase....
QueryDB - ARL.CloudComputing.EntityBase.Entitybase.LoadAsGlobalGraph(2) -       EntityBase.LoadAsGlobalGraph : Loading GG cached Person-to-Person Rel
QueryDB - ARL.CloudComputing.EntityBase.Entitybase.LoadAsGlobalGraph(2) -       EntityBase.LoadAsGlobalGraph : Loading GG cached Person-to-Organizat
QueryDB - ARL.CloudComputing.EntityBase.Entitybase.LoadAsGlobalGraph(2) -
Begin loading relationships of each Organization....
```

Figure B-4. View of “stdout” log file with ERW debug print outs.

1 DEFENSE TECHNICAL
(PDF) INFORMATION CTR
DTIC OCA

2 DIRECTOR
(PDFS) US ARMY RESEARCH LAB
RDRL CIO LT
IMAL HRA MAIL & RECORDS MGMT

1 DIRECTOR
(PDF) US ARMY RESEARCH LAB
RDRL-CII-B M LEE

5 US ARMY RESEARCH LAB
(PDFS) ATTN RDECOM CERDEC D DUFF
ATTN RDER IWP A HANSEN
ATTN RDRL CII C M THOMAS
ATTN RDRL CII C M MITTRICK
ATTN RDRL CII C J RICHARDSON

5 US ARMY RESEARCH LAB
(PDFS) ATTN RDRL CII B BROOME
ATTN RDRL CII B L TOKARCIK
ATTN RDRL CII B R WINKLER
ATTN RDRL CII T V HOLLAND
ATTN RDRL CII T M VANNI

INTENTIONALLY LEFT BLANK.